Statistica Sinica Preprint No: SS-2020-0081

Title	On Semiparametric Instrumental Variable Estimation of Average Treatment Effects through Data Fusion
Manuscript ID	SS-2020-0081
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0081
Complete List of Authors	BaoLuo Sun and Wang Miao
Corresponding Author	BaoLuo Sun
E-mail	stasb@nus.edu.sg

Notice: Accepted version subject to English editing.

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Statistica Sinica

On Semiparametric Instrumental Variable Estimation of Average Treatment Effects through Data Fusion

BaoLuo Sun* and Wang Miao†

* *Department of Statistics and Applied Probability,*

National University of Singapore

†*Department of Probability and Statistics,*

Peking University

Abstract:

Suppose one is interested in estimating causal effects in the presence of potentially unmeasured confounding with the aid of a valid instrumental variable. This paper investigates the problem of making inferences about the average treatment effect when data are fused from two separate sources, one of which contains information on the treatment and the other contains information on the outcome, while values for the instrument and a vector of baseline covariates are recorded in both. We provide a general set of sufficient conditions under which the average treatment effect is nonparametrically identified from the observed data law induced by data fusion, even when the data are from two heterogeneous populations, and derive the efficiency bound for estimating this causal parameter. For inference, we develop both parametric and semiparametric methods, including

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

a multiply robust and locally efficient estimator that is consistent even under partial misspecification of the observed data model. We illustrate the methods through simulations and an application on public housing projects.

Key words and phrases: Two-sample inference; Multiple robustness; Unmeasured confounding

1. Introduction

The instrumental variable method is widely used in the health and social sciences for identification and estimation of causal effects in the presence of potentially unmeasured confounding (Bowden and Turkington, 1990; Robins, 1994; Angrist et al., 1996; Greenland, 2000; Wooldridge, 2010; Hernán and Robins, 2006; Didelez et al., 2010). A valid instrumental variable Z is a pre-exposure variable that is (a) associated with treatment D, (b) independent of any unmeasured confounder U of the exposure-outcome association, and (c) has no direct causal effect on the outcome Y, conditional on a set of measured baseline covariates X. The instrumental variable approach has a longstanding tradition in econometrics going back to the original works of Wright (1928) and Goldberger (1972) in the context of linear structural modeling; see Wooldridge (2010), Clarke and Windmeijer (2012), Baiocchi et al. (2014) and Swanson et al. (2018) for more recent reviews. Under

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

correct specification of the linear structural equation models and assuming absence of baseline covariates, the conventional instrumental variable estimand of the average treatment effect is the population moment ratio $\mathrm{cov}(Z, Y)/\mathrm{cov}(Z, D)$.

However, in many empirical scenarios only information on (Y, Z, X) is available from the primary population of interest. Angrist and Krueger (1992) and Arellano and Meghir (1992) showed that the two sets of moments can be estimated from two separate sources by leveraging information on (D, Z, X) from an auxiliary population, a method known as two-sample instrumental variable estimation. Furthermore, Klevmarken (1982) and Angrist and Krueger (1995) introduced two-sample two-stage least squares estimation with first stage regression for the treatment model based on the auxiliary sample; see Ridder and Moffitt (2007) and Angrist and Pischke (2008) for reviews. This methodology has since been widely applied in econometrics and social sciences (Inoue and Solon, 2010), and more recently in two-sample Mendelian randomization studies to estimate causal relationships using genetic factors as instruments (Pierce and Burgess, 2013; Gamazon et al., 2015; Lawlor, 2016; Zhao et al., 2018, 2019). As noted by Zhao et al. (2019), the aforementioned methods typically assume that the auxiliary data are also sampled from the primary population. In addition,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

linear structural models impose strong homogeneity assumptions on the treatment effect. A robust analytic framework for instrumental identification and estimation of causal effects under data fusion therefore remains of keen interest in observational studies. Graham et al. (2016) identified the two-sample instrumental variable problem as one specific example of a general class of data combination models, and extended the semiparametric efficiency theory of Hahn (1998) and Chen et al. (2008) to this class of models. Recent work has also made significant strides towards relaxing the assumptions for identification of causal effects under data fusion (Pacini and Windmeijer, 2016; Choi et al., 2018; Zhao et al., 2018; Buchinsky et al., 2018; Shu and Tan, 2019; Zhao et al., 2019; Pacini, 2019).

When full data on $L = (Y, D, Z, X)$ are available from the primary population of interest, Robins (1994), Imbens and Angrist (1994), Angrist et al. (1996) and Heckman (1997) formalized the instrumental variable approach under the potential outcome framework (Neyman, 1923; Rubin, 1974), which allows one to nonparametrically define the causal estimands of interest. In this paper, we propose novel assumptions under which the average treatment effect of D on Y in the primary population of interest can be uniquely and nonparametrically identified from the observed data law induced by data fusion. To estimate this identifying statistical functional,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

we develop a suite of parametric and semiparametric estimators including a multiply robust and locally efficient one that remains consistent even if the observed data model is partially misspecified. We compare the proposed estimators both in theory and via simulations, and investigate issues of efficiency and robustness of existing estimators.

2. Model

Suppose we are interested in estimating the average treatment effect of a binary treatment D on outcome Y in a primary population of interest, which is confounded by measured covariates X as well as unmeasured ones U, with the aid of a binary instrumental variable Z. However, we only observe $\{(Y_i, Z_i, X_i)^T,\ i = 1, ..., n_p\}$ from this population. As a remedy, suppose an additional sample $\{(D_i, Z_i, X_i)^T,\ i = 1, ..., n_a\}$ is available from an auxiliary population, possibly different from the primary population. Similar to Graham et al. (2016); Shu and Tan (2019), we assume the following about the data source mechanism:

Assumption 1 (Binomial sampling)**.** *The combined set of $n = n_p + n_a$ units are independent and drawn from either the primary population with a fixed probability $Q_0 \in (0, 1)$ or the auxiliary population with probability $1 - Q_0$.*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Let R_i be an indicator variable, equal to 1 if the ith unit is drawn from the primary population, and 0 otherwise. By assumption 1, the combined set of observed data $\{O_i = (R_i, R_iY_i, (1-R_i)D_i, Z_i, X_i)^T,\ i = 1, ..., n\}$ can be treated as a random sample from a synthetic merged population. Let $F(O)$ denote the distribution of O, with density with respect to some dominating measure given by

$$\begin{aligned} f(O) = q^{\dagger R}(1-q^{\dagger})^{1-R} f(V|R=1)^R f(V|R=0)^{1-R} \times \\ f(Y|V,R=1)^R f(D|V,R=0)^{1-R}, \end{aligned} \tag{2.1}$$

where $V = (Z, X)$ and $q^{\dagger} = \mathrm{pr}(R = 1)$. Let $E(\cdot)$ denote expectation taken with respect to this mixture distribution, and let $\pi(z, x) = E(R|Z = z, X = x)$. By Bayes' rule,

$$f(z, x|R = 1) = f(z, x|R = 0) \left\{ \frac{1-q^{\dagger}}{q^{\dagger}} \frac{\pi(z, x)}{1-\pi(z, x)} \right\}.$$

Let Y_d for $d \in \{0, 1\}$ denote the potential outcome that would be observed if D were set to d, which is related to the observed data via the consistency assumption $Y_d = Y$ if $D = d$. To achieve identification of $\Delta \equiv E(Y_1 - Y_0|R = 1)$ based on the observed data law $F(O)$ induced by data fusion, we make the following assumptions about the primary and auxiliary populations.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2.1 Primary population

Suppose Z is a valid binary instrument that satisfies the following assumptions (Didelez and Sheehan, 2007; Pearl, 2009; Clarke and Windmeijer, 2012):

Assumption 2 (Instrument Relevance)**.** $Z \not\perp\!\!\!\perp D | X, R = 1.$

Assumption 3 (Instrument Independence)**.** $Z \perp\!\!\!\perp U | X, R = 1.$

Assumption 4 (Exclusion Restriction)**.** $Y \perp\!\!\!\perp Z | D, X, U, R = 1.$

Here $A \perp\!\!\!\perp B | C$ indicates conditional independence of A and B given C (Dawid, 1979). Instrument relevance ensures that Z is a correlate of the exposure even after conditioning on X, while instrument independence states that Z is independent of all unmeasured confounders of the exposure-outcome association. Exclusion restriction formalizes the assumption of no direct effect of Z on Y not mediated by D. Furthermore, the assumption of no unmeasured confounding given (X, U) can be stated as

Assumption 5 (Latent Ignorability)**.** $Y_d \perp\!\!\!\perp D | X, U, R = 1$*, for* $d \in \{0, 1\}$ *(Robins, 1994).*

Assumptions 2–5 may be known to hold at the design stage when the investigator controls treatment allocation conditional on baseline covariates

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

in double blind randomized trials. In observational studies, the potential instrumental variable may be viewed as being randomized through some natural or quasi-experiment within levels of the observed covariates (Hernán and Robins, 2006), although these assumptions are typically untestable without further conditions. The exclusion restriction assumption 4 implies the following semiparametric structural models:

$$\begin{aligned} E(D \mid Z, X, U, R=1) &= g_0(X,U) + g_1(X,U)Z \\ E(Y \mid D, Z, X, U, R=1) &= h_0(X,U) + h_1(X,U)D, \end{aligned} \tag{2.2}$$

where for $k \in \{0,1\}$, $g_k(\cdot)$ and $h_k(\cdot)$ are arbitrary square-integrable functions of (X,U) that are only restricted by natural features of the model, e.g. such that the exposure mean is bounded between zero and one. Note that for binary (Z,D), model (2.2) is saturated as there are no restrictions on the corresponding data laws $f(D|Z,X,U,R=1)$ and $f(Y|D,Z,X,U,R=1)$ except for the implications of assumption 4. Under assumptions 4 and 5, $h_1(x,u) = E(Y_1 - Y_0|X=x, U=u, R=1)$ encodes the conditional average treatment effect within levels of (X,U), hence $\Delta = E\{h_1(X,U)|R=1\}$. The linear structural equation model (Wright, 1928; Goldberger, 1972)

$$\begin{aligned} E(D \mid Z, X, U, R=1) &= \theta_0 + \theta_1 X + \theta_2 U + \theta_3 Z \\ E(Y \mid D, Z, X, U, R=1) &= \beta_0 + \beta_1 X + \beta_2 U + \Delta D, \end{aligned} \tag{2.3}$$

is a special case of (2.2), where the function $h_1(X,U)$ is reduced to the scalar

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

parameter of interest Δ encoding the homogeneous average treatment effect within levels of (X, U).

Even when full data on $L = (Y, D, Z, X)$ are available from the primary population, it is well known that while a valid instrumental variable satisfying assumptions 2–5 suffices to obtain a valid statistical test of the sharp null hypothesis of no individual causal effect, the population average treatment effect Δ is itself not uniquely identified from the law $F(L|R = 1)$ (Balke and Pearl, 1997). With a further monotonicity assumption about the effect of Z on D, Angrist et al. (1996) showed that the local average treatment effect (LATE) among compliers can be nonparametrically identified. This framework has been further generalized in recent years by Abadie et al. (2002), Abadie (2003), Carneiro et al. (2003), Tan (2010a), Ogburn et al. (2015) and Kennedy et al. (2019). Zhao et al. (2019) discussed identification of LATE in two-sample instrumental variable analyses. However, because the population of compliers is itself nonidentifiable in general, Δ is arguably still a causal parameter of interest in many observational studies (Robins and Greenland, 1996; Imbens, 2010). Wang and Tchetgen Tchetgen (2018) proved identifiability of Δ from the law $F(L|R = 1)$ under the additional assumption

$$g_1(X, U) = g_1(X) \quad \text{or} \quad h_1(X, U) = h_1(X) \quad \text{almost surely,} \tag{2.4}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

i.e. at least one of these effects is not allowed to vary with U. We show that Δ can be identified from $F(O)$ provided X must be sufficiently rich so that the effect of exposure on the outcome is uncorrelated with the effect of the instrument on the exposure conditional on X (Cui and Tchetgen Tchetgen, 2019), which can be achieved even if X does not include all confounders of the effect of D on Y.

Assumption 6 (Orthogonality). *$cov\{g_1(X,U), h_1(X,U)|X, R = 1\} = 0$ almost surely.*

Assumption 6 may hold under certain data generating mechanisms even if (2.4) does not, and is guaranteed to hold under the sharp causal null effect. In addition, we require every unit within levels of the observed covariates to have some chance of receiving each level $z \in \{0,1\}$ of the instrument.

Assumption 7 (Positivity). *$0 < pr(Z = 1|X, R = 1) < 1$ almost surely.*

2.2 Auxiliary population

We make the following assumptions about the auxiliary population:

Assumption 8 (Support overlap). *$0 < \pi(Z, X) < 1$ almost surely.*

Assumption 9 (Propensity score equality). *$pr(D = 1|Z, X, R = 0) = pr(D = 1|Z, X, R = 1)$ almost surely.*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Assumption 8 ensures that the support of the common variables (Z, X) in the primary population is contained within that in the auxiliary population, and together with assumption 9 allows us to identify the treatment propensity score $\tau(z, x) = \text{pr}(D = 1|Z = z, X = x, R = 1)$ based on $F(O)$. Assumption 9 only requires predictive invariance for the treatment between the two heterogeneous populations, and we do not require the stronger condition of "structural invariance" (e.g. assumptions 3–6 also hold in the auxiliary population), which is related to the notions of "invariant prediction" (Peters et al., 2016), "autonomy" (Haavelmo, 1944) and "stability" (Pearl, 2009) as discussed in Zhao et al. (2019).

2.3 Nonparametric identification

We show that under assumptions 1–9, Δ is a functional on the nonparametric observed data statistical model $\mathcal{M}_{\text{np}} = \{F(O) : F(O) \text{ unrestricted}\}$ of all regular laws $F(O)$ that satisfy the positivity and support overlap assumptions. In the following, let $\lambda(z|x) = \text{pr}(Z = z|X = x, R = 1)$ denote the probability density or mass function of Z given X in the primary population.

Theorem 1. *Under assumptions 1–9,*

$$\Delta = E\left\{\frac{R}{q^{\dagger}}\frac{(-1)^{1-Z}}{\lambda(Z|X)}\frac{Y}{[\tau(1, X) - \tau(0, X)]}\right\}. \tag{2.5}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Remark 1. When Y is continuous and D and Z are discrete of finite domain, the canonical instrumental variable assumptions 3 and 4 impose no constraints on the law $F(L|R=1)$ (Bonet, 2001). In addition, assumption 9 is akin to coarsening at random, which leaves the observed data law $F(O)$ unrestricted (Robins, 1997; Van der Laan et al., 2003). When Y is also discrete, assumptions 3 and 4 impose inequality constraints which do not restrict the parameter space of $F(L|R=1)$ locally if the true observed data law lies in the interior of the space defined by these constraints (Wang et al., 2017; Wang and Tchetgen Tchetgen, 2018).

Remark 2. While nuisance parameters such as $\{\lambda(\cdot), \tau(\cdot)\}$ can in principle be estimated nonparametrically using methods such as sieve estimation (Hahn, 1998; Hirano et al., 2003; Chen et al., 2008), in this paper we focus on parametric working models due to the curse of dimensionality when X is of moderate or high dimension (Robins and Ritov, 1997). Since one cannot be confident that any of these models is correctly specified, we also propose an estimator of Δ that is robust to misspecifications of these models.

Remark 3. The form of the identification formula (2.5) in Theorem 1 suggests that Δ may be identified as long as one has access to consistent estimators of the propensity score $\tau(z, x)$ in the primary population. The utility of the sample from the auxiliary population lies in estimation of

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\tau(z,x)$ under Assumption 9, which is not testable as D is not observed in the sample from the primary population. On the other hand, Δ may be identified without the need for an auxiliary sample if the propensity score is known by design in the primary population.

3. Estimation

3.1 Maximum likelihood estimation

Let $\hat{E}(\cdot)$ denote the empirical mean operator $\hat{E}\{h(O)\} = n^{-1}\sum_{i=1}^{n} h(O_i)$, and let $(\hat{\alpha}, \hat{\psi}, \hat{\xi}, \hat{\theta})$ denote the maximum likelihood estimators of $(\alpha, \psi, \xi, \theta)$ that index the parametric models $\pi(z, x; \alpha)$, $\lambda(z|x; \psi)$, $\tau(z, x; \xi)$ and additionally $f(y|z, x, R = 1; \theta) = f(Y = y|Z = z, X = x, R = 1; \theta)$ for the outcome conditional density specified by the analyst. We note that under assumption 9, $\tau(z, x) = \mathrm{pr}(D = 1|Z = z, X = x, R = 0)$ so that inferences on ξ can be based on the auxiliary sample. By taking iterated expectation of (2.5) with respect to (Z, X), the plug-in estimator of Δ is

$$\hat{\Delta}_{\mathrm{mle}} = \hat{E}\left\{\frac{1}{\hat{q}}\frac{(-1)^{1-Z}}{\lambda(Z|X;\hat{\psi})}\frac{\pi(Z,X;\hat{\alpha})E(Y|Z,X,R=1;\hat{\theta})}{\tau(1,X;\hat{\xi})-\tau(0,X;\hat{\xi})}\right\}, \tag{3.6}$$

where the distribution of (Z, X) is estimated by its empirical distribution and $\hat{q} = \hat{E}(R)$. It is clear that consistency of $\hat{\Delta}_{\mathrm{mle}}$ relies on correct specifications of the models $\pi(z, x; \alpha)$, $\lambda(z|x; \psi)$, $\tau(z, x; \xi)$ and $f(y|z, x, R = 1; \theta)$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In the following we propose several semiparametric estimators of Δ that do not require these models to be fully specified. We proceed by first noting the following decomposition of the outcome conditional mean model.

Lemma 1. *Under assumptions 2–6,*

$$E(Y|Z=z, X=x, R=1) = \mathcal{H}(x)\tau(z,x) + \omega(x), \tag{3.7}$$

where $\omega(x) \equiv cov[g_1(X,U), h_1(X,U)|X=x, R=1] + E[h_0(X,U)|X=x, R=1]$ *and* $\mathcal{H}(x) \equiv E[h_1(U,X)|X=x, R=1]$ *is the treatment effect curve conditional on observed covariates. Therefore,* $\Delta = E\{\mathcal{H}(X)|R=1\}$.

3.2 Semiparametric estimation

Consider the following submodels of $\mathcal{M}_{\text{np}}$ in which smooth parametric models (indexed by finite-dimensional parameters) for certain components of the observed data law $F(O)$ are correctly specified:

Definition 1.

$\mathcal{M}_1$: The models $\lambda(z|x;\psi)$ and $\tau(z,x;\xi)$ are correctly specified such that $\lambda(z|x;\psi^\dagger) = \lambda(z|x)$ and $\tau(z,x;\xi^\dagger) = \tau(z,x)$ for some unknown values $(\psi^\dagger, \xi^\dagger)$;

$\mathcal{M}_2$: The models $\mathcal{H}(x;\gamma)$, $\omega(x;\eta)$ and $\tau(z,x;\xi)$ are correctly specified such that $\mathcal{H}(x;\gamma^\dagger) = \mathcal{H}(x)$, $\omega(x;\eta^\dagger) = \omega(x)$ and $\tau(z,x;\xi^\dagger) = \tau(z,x)$ for some

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

unknown values $(\gamma^\dagger, \eta^\dagger, \xi^\dagger)$;

$\mathcal{M}_3$: The models $\mathcal{H}(x;\gamma)$, $\omega(x;\eta)$ and $\pi(z,x;\alpha)$ are correctly specified such that $\mathcal{H}(x;\gamma^\dagger) = \mathcal{H}(x)$, $\omega(x;\eta^\dagger) = \omega(x)$ and $\pi(z,x;\alpha^\dagger) = \pi(z,x)$ for some unknown values $(\gamma^\dagger, \eta^\dagger, \alpha^\dagger)$.

We propose semiparametric estimators for Δ which are consistent and asymptotically normal in each of the above submodels. Our first estimator $\hat{\Delta}_1$ of Δ is motivated by identification formula (2.5) which does not require specification of an outcome model for $f(y|z,x,R=1)$, and solves

$$\begin{aligned} 0 &= \hat{E}\left\{\mu_1(O;\Delta,\hat{\psi},\hat{\xi},\hat{q})\right\} \\ &\equiv \hat{E}\left\{\frac{R}{\hat{q}}\frac{(-1)^{1-Z}}{\lambda(Z|X;\hat{\psi})}\frac{Y}{[\tau(1,X;\hat{\xi})-\tau(0,X;\hat{\xi})]} - \Delta\right\}. \end{aligned} \tag{3.8}$$

Remark 4. The models for $\{\lambda(\cdot),\tau(\cdot)\}$ can be specified and estimated without access to the outcome data. Estimation of Δ using $\hat{\Delta}_1$ could therefore be considered as part of a more objective analysis design in the sense that it mitigates potential for "data-dredging" exercises when the outcome model is fully specified (Rubin, 2007).

We propose two additional estimators of Δ which do not require a model for $\lambda(\cdot)$ but instead posit models $\mathcal{H}(X;\gamma)$ and $\omega(X;\eta)$ for components of the outcome conditional mean (3.7). Consider the semiparametric estimators

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\hat{\Delta}_2$ and $\hat{\Delta}_3$ which solve

$$0 = \hat{E}\left\{\mu_j(O; \Delta, \hat{\gamma}_j, \hat{q})\right\} \equiv \hat{E}\left\{\frac{R}{\hat{q}}[\mathcal{H}(X; \hat{\gamma}_j) - \Delta]\right\} \tag{3.9}$$

for $j = 1, 2$ respectively, where the estimators $\hat{\gamma}_2$ and $\hat{\gamma}_3$ are constructed in a way such that they are consistent in the submodels $\mathcal{M}_2$ and $\mathcal{M}_3$ respectively, as follows. Let $v(X)$ and $w(X)$ be analyst-specified vector functions of the same dimensions as γ and η respectively, for example $\{v(X), w(X)\} = \{\partial\mathcal{H}(X;\gamma)/\partial\gamma, \partial\omega(X;\eta)/\partial\eta\}$, and let $\mathcal{G}_{v,w}(X, Z) = \{v^T(X)Z, w^T(X)\}^T$ where A^T denotes the transpose of A. Then let $(\hat{\gamma}_2, \hat{\eta}_2)$ be the joint solution to the estimating equation

$$\begin{aligned} 0 &= \hat{E}\big\{\mathcal{G}_{v,w}(X, Z)\big\{R[Y - \mathcal{H}(X;\gamma)\tau(Z, X; \hat{\xi}) - \omega(X;\eta)] \\ &\quad -(1-R)\mathcal{H}(X;\gamma)[D - \tau(Z, X; \hat{\xi})]\big\}\big\}, \end{aligned}$$

while $(\hat{\gamma}_3, \hat{\eta}_3)$ jointly solve

$$0 = \hat{E}\left\{\mathcal{G}_{v,w}(X, Z)\left\{R[Y - \omega(X;\eta)] - \frac{(1-R)\pi(Z, X; \hat{\alpha})}{1 - \pi(Z, X; \hat{\alpha})}\mathcal{H}(X;\gamma)D\right\}\right\}.$$

Lemma 2. *Under standard regularity conditions (Newey and McFadden, 1994), the estimators $\hat{\Delta}_1$, $\hat{\Delta}_2$, and $\hat{\Delta}_3$ are consistent and asymptotically normal in submodels $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, respectively.*

Remark 5. To ensure that the proposed estimators of Δ lie between -1 and 1 in the case of binary Y, following Wang and Tchetgen Tchetgen

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

(2018) we can specify a model such as

$$\mathcal{H}(X;\gamma) = \tanh(\gamma^T X) = \frac{\exp(2\gamma^T X) - 1}{\exp(2\gamma^T X) + 1},$$

which guarantees that $\mathcal{H}(X;\gamma) \in [-1, 1]$. In addition, instead of the decomposition (3.7) for continuous Y, Wang and Tchetgen Tchetgen (2018) provided a variation independent decomposition of the components in the likelihood $\{\mathrm{pr}(Y = 1|Z, X, R = 1), \mathrm{pr}(D = 1|Z, X, R = 1)\}$ for binary Y, and their estimation strategy for these components may be adopted similarly.

3.3 Multiply robust estimation

To motivate the multiply robust estimator, we consider efficient estimation of Δ in $\mathcal{M}_{\text{np}}$. Any regular and asymptotically linear estimator $\hat{\Delta}$ has an associated influence function $\mu(O;\Delta)$ such that $\hat{\Delta} - \Delta = \hat{E}\{\mu(O;\Delta)\} + o_p(n^{-1/2})$ (Bickel et al., 1993). Therefore it suffices to identify $\mu(O;\Delta)$ with the lowest variance, which is the efficient influence function.

Theorem 2. *The efficient influence function for Δ in $\mathcal{M}_{np}$ is*

$$\mu_{\text{eff}}(O;\Delta) = \frac{(-1)^{1-Z}\left\{\begin{array}{c}\frac{R}{q^\dagger}[Y - \mathcal{H}(X)\tau(Z,X) - \omega(X)] \\ -\frac{1-R}{q^\dagger}\frac{\pi(Z,X)}{1-\pi(Z,X)}\mathcal{H}(X)[D - \tau(Z,X)]\end{array}\right\}}{\lambda(Z|X)[\tau(1,X) - \tau(0,X)]} + \frac{R}{q^\dagger}\{\mathcal{H}(X) - \Delta\}, \tag{3.10}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

so that the semiparametric efficiency bound for estimating Δ *in* $\mathcal{M}_{np}$ *is* $E\{\mu^2_{\text{eff}}(O;\Delta)\}$.

We use $\mu_{\text{eff}}(\cdot)$ as an estimating function and plug in estimates of the nuisance parameters to estimate the causal effect Δ. This method of constructing estimating equations from influence functions has been used widely, e.g. in Bang and Robins (2005); Tan (2006b); Tchetgen Tchetgen et al. (2009); Sun et al. (2018); Sun and Tchetgen Tchetgen (2018); Wang and Tchetgen Tchetgen (2018). Consider $(\tilde{\gamma},\tilde{\eta})$ which jointly solve

$$\begin{aligned}\mathbf{0} = \ & \hat{E}\Bigg\{\mathcal{G}_{v,w}(X,Z)\Bigg\{R[Y-\mathcal{H}(X;\gamma)\tau(Z,X;\hat{\xi})-\omega(X;\eta)] \\ & -\frac{(1-R)\pi(Z,X;\hat{\alpha})}{1-\pi(Z,X;\hat{\alpha})}\mathcal{H}(X;\gamma)[D-\tau(Z,X;\hat{\xi})]\Bigg\}\Bigg\}.\end{aligned} \tag{3.11}$$

We note that the estimator $\tilde{\gamma}$ is doubly robust in the sense that it is consistent for $\gamma^\dagger$ in the model $\mathcal{M}_2\cup\mathcal{M}_3$, which is necessary for the multiply robust result stated below.

Lemma 3. *Under standard regularity conditions (Newey and McFadden, 1994), the estimator* $\hat{\Delta}_{\text{mul}}$ *which solves*

$$0 \ = \ \hat{E}\left\{\mu_{\text{eff}}(O;\Delta,\tilde{\eta},\tilde{\gamma},\hat{\psi},\hat{\xi},\hat{\alpha},\hat{q})\right\} \tag{3.12}$$

is consistent and asymptotically normal in the union model $\mathcal{M}_{\text{union}}=\cup_{j=1}^3\mathcal{M}_j$ *(multiply robust). Moreover,* $\hat{\Delta}_{\text{mul}}$ *attains the semiparametric efficiency*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

bound in $\mathcal{M}_{np}$ (and, following the general results of Robins and Rotnitzky (2001), also in $\mathcal{M}_{\text{union}}$) at the intersection submodel $\{\cap_{j=1}^{3}\mathcal{M}_j\}$ where all working models are correctly specified (locally efficient).

The asymptotic variance formula of each estimator described in this section follows from standard M-estimation theory (Newey and McFadden, 1994). For inference based on the proposed semiparametric estimators of Δ in both the simulation study and application (sections 5 and 6 respectively), consistent estimation of the asymptotic variance is described in the Supplementary Materials.

4. Comparison to some existing estimators

Suppose that $E(U|Z=z,X=x,R=1)=E(U|X=x,R=1)$ is linear in x, then the linear structural models (2.3) yield the observed data models

$$\begin{aligned}
\tau_{\text{linear}}(Z,X;\xi) &= \xi^T(1,Z,X)^T;\\
\omega_{\text{linear}}(X;\eta) &= \eta^T(1,X)^T;\\
E(Y\mid Z,X,R=1) &= \Delta\tau_{\text{linear}}(Z,X;\xi)+\omega_{\text{linear}}(X;\eta).
\end{aligned}$$

We also have that $\mathcal{H}(X)$ is indexed by the scalar parameter of interest Δ. Using the notation in section 3, it can be shown that the two-sample

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

instrumental variable estimator (Inoue and Solon, 2010) $(\hat{\Delta}_{\text{tsiv}}, \hat{\eta}_{\text{tsiv}})$ solves

$$0 = \hat{E}\Big\{\mathcal{G}_{v,w}(X,Z)\Big\{R[Y - \omega_{\text{linear}}(X;\eta)] - \frac{(1-R)\hat{q}}{1-\hat{q}}\Delta D\Big\}\Big\}.$$

Inferences based on the two-sample instrumental variable estimator can be viewed as special instances of inferences obtained under a particular specification of submodel $\mathcal{M}_3$ with the above parametric models for $\{\mathcal{H}(\cdot), \omega(\cdot)\}$ and additionally $\pi(z,x;\alpha) = q$ where $q \in \mathbb{R}$, e.g. the marginal distribution of (Z,X) is the same in the primary and auxiliary populations. Therefore $\hat{\Delta}_{\text{tsiv}}$ will fail to be consistent for Δ if any of the parametric models in $\mathcal{M}_3$ is incorrectly specified. Furthermore, we note that the two-sample two-stage least squares estimator $(\hat{\Delta}_{\text{ts2sls}}, \hat{\eta}_{\text{ts2sls}})$ solves

$$\begin{aligned} 0 &= \hat{E}\Big\{\mathcal{G}_{v,w}(X,Z)\Big\{R[Y - \Delta\tau_{\text{linear}}(Z,X;\hat{\xi}) - \omega_{\text{linear}}(X;\eta)] \\ &- \frac{(1-R)\hat{q}}{1-\hat{q}}\Delta[D - \tau_{\text{linear}}(Z,X;\hat{\xi})]\Big\}\Big\}, \end{aligned}$$

which is a special case of the doubly robust estimating equation (3.11). It follows that $\hat{\Delta}_{\text{ts2sls}}$ is consistent for Δ in $\mathcal{M}_2 \cup \mathcal{M}_3$; even when the true marginal distribution of (Z,X) differs between the primary and auxiliary populations, $\hat{\Delta}_{\text{ts2sls}}$ is consistent provided the linear propensity score model $\tau_{\text{linear}}(\cdot)$ is correctly specified. We can also show via semiparametric effciency theory that $\hat{\Delta}_{\text{ts2sls}}$ is asymptotically more efficient than its non-doubly robust counterpart $\hat{\Delta}_{\text{tsiv}}$ at the intersection submodel $\mathcal{M}_2 \cap \mathcal{M}_3$ (Tan, 2007;

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Tsiatis, 2007). The above properties were noted by Inoue and Solon (2010).

Shu and Tan (2019) proposed a class of doubly robust estimators $(\hat{\Delta}_{\rm dr}, \hat{\eta}_{\rm dr})^T$ which solve

$$\begin{aligned} 0 &= \hat{E}\Big\{\mathcal{G}_{v,w}(X,Z)\Big\{R[Y-\Delta\tau(Z,X;\hat{\xi})-\omega_{\rm linear}(X;\eta)] \\ &- \frac{(1-R)\pi(Z,X;\hat{\alpha})}{1-\pi(Z,X;\hat{\alpha})}\Delta[D-\tau(Z,X;\hat{\xi})]\Big\}\Big\}, \end{aligned}$$

where users can freely specify models for $\{\tau(\cdot),\pi(\cdot)\}$. Graham et al. (2016) introduced in earlier work a doubly robust auxiliary-to-study tilting estimator under restricted nuisance model specifications in efficient estimation of data combination models. Inferences based on $\hat{\Delta}_{\rm dr}$ can be viewed as special instances of inferences obtained under a particular specification of submodel $\mathcal{M}_2 \cup \mathcal{M}_3$ with $\mathcal{H}(X) = \Delta$ and $\omega_{\rm linear}(\cdot)$. In constrast to $\hat{\Delta}_{\rm mul}$, $\hat{\Delta}_{\rm dr}$ will generally fail to be consistent for Δ outside the union model $\mathcal{M}_2 \cup \mathcal{M}_3$. We note that a generalized version of $\hat{\Delta}_{\rm dr}$ that accommodates arbitrary parametric model specifications in $\mathcal{M}_2 \cup \mathcal{M}_3$ is given by

$$\hat{\Delta}_{\rm dr2} = \hat{E}\left\{R\mathcal{H}(X;\tilde{\gamma})/\hat{q}\right\}, \tag{4.13}$$

where $\tilde{\gamma}$ solves (3.11).

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

5. Simulation study

We investigate the finite-sample properties of the proposed semiparametric estimators under a variety of settings. For the primary population, baseline covariates $X = (X_1, X_2, X_3)^T$ are mutually independent and marginally distributed as U$(0, 1)$; (Y, A, Z, U) is distributed as follows:

$$
\begin{aligned}
U|X &\sim \mathrm{TN}\{\vartheta^T X, 1, (\vartheta^T X - 1, \vartheta^T X + 1)\};\\
Z|X &\sim \text{Bernoulli } \{p = \{1 + \exp[-\psi^T (1, X^T)^T]\}^{-1}\};\\
D|Z, X, U &\sim \text{Bernoulli } \{p = \{1 + \exp[-\xi^T (1, Z, X^T)^T]\}^{-1} + 0.2[U - \vartheta^T X]\};\\
Y|D, X, U &\sim \mathrm{N}\{\gamma^T (1, X^T)^T D + 1.25 \times \vec{1}^T X + 6U, 1\},
\end{aligned}
$$

where $\mathrm{TN}\{\mu, \sigma^2, (l, u)\}$ denotes a truncated normal distribution with support $[l, u]$, $\vartheta = (0.5, -0.5, 0)^T$, $\psi = (-1, 0.5, 0.5, 0.5)^T$, $\xi = (-1.3, 1.2, 0.5, -0.25- 0.25)^T$, $\gamma = (2, 0.5, 0.5, 0.5)^T$ and $\vec{1} = (1, 1, 1)^T$. For the auxiliary population, $X = (X_1, X_2, X_3)^T$ are mutually independent and marginally distributed as $\mathrm{TN}\{0.5, 1, (0, 1)\}$, $Z|X \sim$ Bernoulli $\{p = \{1+\exp[-\psi^T (1, X^T)^T]\}^{-1}\}$ and $D|Z, X \sim$ Bernoulli $\{p = \{1+\exp[-\xi^T (1, Z, X^T)^T]\}^{-1}\}$; the remaining parts of the data law are left unrestricted. For each simulation replicate of total sample size n, we generate $n_p \sim \text{binomial}(n, p = 0.7)$, followed by an i.i.d. sample of size n_p from the primary population with only realizations of (Y, Z, X) recorded, and another i.i.d. sample of size $n_a = n - n_p$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

from the auxiliary population with only realizations of (D, Z, X) recorded. The two samples are then merged, and an indicator variable R is introduced, equal to 1 or 0 if the unit is drawn from the primary or auxiliary population respectively. It can be verified that the above data generating mechanism satisfies assumptions 1–9, and that the corresponding true observed data models are $\lambda(1|x;\psi) = \{1 + \exp[-\psi^T(1, x^T)^T]\}^{-1}$, $\tau(z, x; \xi) = \{1 + \exp[-\xi^T(1, z, x^T)^T]\}^{-1}$, $\mathcal{H}(x;\gamma) = \gamma^T(1, x^T)^T$, $\omega(x;\eta) = \eta^T(1, x^T)^T$ and $\pi(z, x; \alpha) = \{1 + \exp[-\alpha^T(1, z, x^T, x^{2T})^T]\}^{-1}$ where $x^2 = (x_1^2, x_2^2, x_3^2)^T$ (by Bayes' rule). We are interested in estimating the average treatment effect $\Delta = E\{\gamma^T(1, X^T)^T|R = 1\} = 2.75$. The four semiparametric estimators $\hat{\Delta}_1$, $\hat{\Delta}_2$, $\hat{\Delta}_3$ and $\hat{\Delta}_{\text{mul}}$ are implemented using $v(x) = w(x) = (1, x^T)^T$ as index functions.

Similar to Kang et al. (2007), we evaluate the performance of the proposed estimators in situations where some models may be mis-specified by considering the transformed variables $V^* = (Z^*, X_1^*, X_2^*, X_3^*)^T$ where $Z^* \sim \text{Bernoulli}\{p = \Phi(-2 + 3Z)\}$, $X_1^* = \exp(-0.5X_1) + \epsilon_1$, $X_2^* = X_2/[1 + \exp(Z)] + \epsilon_2$ and $X_3^* = (X_1X_3)^3 + \epsilon_3$; $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and the error terms are generated as $(\epsilon_1, \epsilon_2, \epsilon_3)^T \sim N(0, I_3)$. Then a particular component model is mis-specified when the analyst uses V^* instead of V in the working model.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Specifically, we report results from the following five scenarios:

$\mathcal{M}'_0$**:** All models are correct;

$\mathcal{M}'_1$**:** Only models $\lambda(z|x;\psi)$ and $\tau(z,x;\xi)$ are correct;

$\mathcal{M}'_2$**:** Only models $\tau(z,x;\xi)$, $\mathcal{H}(x;\gamma)$ and $\omega(x;\eta)$ are correct;

$\mathcal{M}'_3$**:** Only models $\pi(z,x;\alpha)$, $\mathcal{H}(x;\gamma)$ and $\omega(x;\eta)$ are correct;

$\mathcal{M}'_4$**:** All models are incorrect.

All simulation results are based on 1000 Monte Carlo runs of $n = 10000$ units each. Table 1 summarizes simulation results. In agreement with theory, $\hat{\Delta}_1$ has small bias in $\mathcal{M}'_0$ and $\mathcal{M}'_1$, $\hat{\Delta}_2$ has small bias in $\mathcal{M}'_0$ and $\mathcal{M}'_2$, $\hat{\Delta}_3$ has small bias in $\mathcal{M}'_0$ and $\mathcal{M}'_3$, and $\hat{\Delta}_{\text{mul}}$ has small bias in $\mathcal{M}'_l$, $l = 0, 1, 2, 3$. In $\mathcal{M}'_0$ where all models are correct, $\hat{\Delta}_1$ and $\hat{\Delta}_2$ have smaller Monte Carlo standard errors compared to $\hat{\Delta}_3$ which involves weighting through the data source propensity score $\pi(z,x)$.

6. Application

Currie and Yelowitz (2000) studied the effect of public housing participation on housing quality and educational attainment, and showed that project participation is associated with poorer outcomes based on data

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 1: Monte Carlo results of the proposed semiparametric estimators under different scenarios

Model	Estimator			
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_{\mathrm{mul}}$
|Bias| (SE)				
$\mathcal{M}_0'$	0.01 (0.29)	0.01 (0.29)	0.08 (0.33)	0.04 (0.31)
$\mathcal{M}_1'$	0.01 (0.29)	0.65 (0.34)	0.74 (0.37)	0.05 (0.30)
$\mathcal{M}_2'$	0.67 (0.36)	0.01 (0.32)	0.11 (0.41)	0.05 (0.33)
$\mathcal{M}_3'$	1.10 (0.46)	1.20 (0.48)	0.09 (0.34)	0.06 (0.33)
$\mathcal{M}_4'$	1.30 (0.47)	2.20 (0.57)	0.77 (0.44)	0.72 (0.39)
RMSE				
$\mathcal{M}_0'$	0.09	0.09	0.11	0.10
$\mathcal{M}_1'$	0.08	0.54	0.68	0.09
$\mathcal{M}_2'$	0.58	0.10	0.18	0.11
$\mathcal{M}_3'$	1.50	1.70	0.12	0.11
$\mathcal{M}_4'$	1.80	5.00	0.78	0.67

from the Survey of Income and Program Participation (SIPP). However, many unobserved factors such as social ties are likely to affect both project participation and outcomes, and the authors suspect that failure to control for this source of endogeneity would bias the estimated causal effects of living in projects downwards, since families in projects may be more likely to live in substandard housing in any case, and their children may be more likely to experience negative outcomes. Leveraging on the sex composition of children as an instrumental variable for project participation, Currie and Yelowitz (2000) use two-sample instrumental variable methods to combine

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

information from the 1990 Census data and 1990-1995 waves of the March Current Population Survey (CPS), and find that project households are less likely to suffer from overcrowding or live in high-density complexes, and project children are less likely to have been held back. Their study is important as the results overturn the stereotype that project participation is harmful in terms of living conditions and children's educational attainment.

In this analysis, we apply the proposed methods to estimate the causal effect of project participation (D) on reported monthly rental payments (Y) in the SIPP population; reported rent may be viewed as a proxy for housing quality (Currie and Yelowitz, 2000). The binary instrumental variable Z takes on value 1 if a family had a boy and a girl, and 0 if both are boys or girls. Families with two children of opposite genders will be eligible for three-bedroom apartments as opposed to two-bedroom apartments, and therefore will be more likely to participate in the housing project, although there is little reason to expect that the children's sex composition will directly affect Y. In line with the Currie and Yelowitz (2000) study, the vector of baseline covariates X includes the household head's gender, age, race, education, marital status and the number of boys in the family. We specify main effects models for $\{\lambda(\cdot), \tau(\cdot), \pi(\cdot)\}$ with logistic links. In addition, fol-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 2: Estimates of the effect of public housing project participation on reported monthly rental (divided by 1000 US dollars).

	point estimate	standard error	95% Wald CI
$\hat{\Delta}_{\mathrm{ts2sls}}$	0.3717	0.1124	(0.1513, 0.5920)
$\hat{\Delta}_1$	0.7650	0.3442	(0.0903, 1.4397)
$\hat{\Delta}_2$	0.3790	0.1162	(0.1513, 0.6068)
$\hat{\Delta}_3$	0.4999	0.2533	(0.0034, 0.9964)
$\hat{\Delta}_{\mathrm{mul}}$	0.9155	0.4126	(0.1069, 1.7242)

lowing Shu and Tan (2019) we add an additional interaction term involving household head information to the linear predictor function of the model for $\pi(\cdot)$ to improve covariate balance, and specify $\omega(x;\eta) = \eta^T(1,x^T)^T$, $\mathcal{H}(x;\gamma) = \Delta$. The analysis results based on $n_1 = 116901$ renters' complete records for (Y,Z,X) from the 1990 Census of SIPP ($R = 1$) and $n_0 = 10382$ renters' complete records for (D,Z,X) from CPS ($R = 0$), for a total sample size of $n = 127283$, are summarized in Table 2.

The two-sample two-stage least squares estimate of 0.3717 agrees with the point estimate presented in Table 4 of Currie and Yelowitz (2000), although the analytic standard error of 0.1124 is larger than the value of 0.0589 reported by the original study, as the former takes into account the variability associated with the first-stage estimation. While the point estimates of the proposed estimators are all larger than 0.3717, the point estimate of $\hat{\Delta}_{\mathrm{mul}}$ is closest to that of $\hat{\Delta}_1$, which suggests that the models

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

for $\{\lambda(\cdot), \tau(\cdot)\}$ in this illustrative analysis may be specified nearly correctly; Tchetgen Tchetgen and Robins (2010) describe a formal specification test to detect which of the baseline models is correct under the union model $\mathcal{M}_{\text{union}}$. The point estimate of 0.9155 for $\hat{\Delta}_{\text{mul}}$ also suggests that the causal effect of housing project participation on improving household living conditions is probably larger than the value reported in Currie and Yelowitz (2000), since $\hat{\Delta}_{\text{ts2sls}}$ is generally no longer consistent outside the union model $\mathcal{M}_2 \cup \mathcal{M}_3$.

7. Discussion

Suppose we observe data on (D, Z, X) from the primary population of interest and fuse it with data on (Y, Z, X) from an auxiliary source, i.e. R_i equals to either 0 or 1 if the ith unit is drawn from the primary or the auxiliary population respectively. In this case, it is clear that inference about the identifying functional

$$\Delta = E\left\{\frac{1-R}{1-q^{\dagger}} \frac{(-1)^{1-Z}}{\lambda(Z|X)} \frac{Y}{[\tau(1,X)-\tau(0,X)]}\right\}$$

is not possible under submodel $\mathcal{M}_1$, since Y is not observed from the primary population. Nonetheless, inference for Δ is still possible under $\mathcal{M}_2 \cup \mathcal{M}_3$ if we replace assumption 9 with predictive invariance for the outcome:

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Assumption 10. $E(Y|Z,X,R=0)=E(Y|Z,X,R=1)$ *almost surely.*

Indeed, it can be shown that under assumptions 1–8 and 10, the estimator

$$\tilde{\Delta}_{\mathrm{dr3}}=\hat{E}\left\{(1-R)\mathcal{H}(X;\tilde{\gamma})/(1-\hat{q})\right\}, \tag{7.14}$$

where $\tilde{\gamma}$ solving (3.11) is consistent and asymptotically normal in the union model $\mathcal{M}_2\cup\mathcal{M}_3$. We note that because $\hat{\Delta}_{\mathrm{tsiv}}$, $\hat{\Delta}_{\mathrm{ts2sls}}$ and $\hat{\Delta}_{\mathrm{dr}}$ typically specify $\mathcal{H}(x;\gamma)=\Delta$ which does not depend on values for the baseline covariates, one can be agnostic as to which of the two samples is drawn from the primary population as long as assumptions 1–10 all hold.

There are several improvements and extensions for future work. Multiple valid instrumental variables can be incorporated by adopting a standard generalized method of moments approach (Hansen, 1982), and the proposed estimators can be improved in terms of efficiency (Tan, 2006a, 2010b) and bias (Vermeulen and Vansteelandt, 2015). In this paper, we focused on the canonical case of binary Z and D; extension of the proposed methodology to the case of general Z or D is an interesting topic for future research. It will also be of interest to investigate the use of negative controls under data fusion to mitigate unmeasured confounding and identify causal effects, which has gained increasing recognition and popularity in recent years (Miao and Tchetgen Tchetgen, 2017; Shi et al., 2018).

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Multiply robust estimation typically entails postulating various parametric models for the nuisance parameters (Molina et al., 2017). Chernozhukov et al. (2018) showed that $n^{-1/2}$ consistent estimation of low-dimensional parameters of interest based on nonparametric efficient scores such as $\mu_{\text{eff}}(O;\Delta)$ is possible when all the nuisance parameters are consistently estimated with sufficiently fast rates, even when the complexity of the nuisance model space is no longer limited by classical settings. In future research, we plan to investigate estimation and inference for the average treatment effect under data fusion when various flexible and highly data-adaptive machine learning methods are used to estimate the nuisance parameters.

Supplementary Materials

Supplementary material available includes the proof of Lemmas and Theorems as well as details on asymptotic variance estimation for the proposed estimators.

Acknowledgments

BaoLuo Sun is supported by the National University of Singapore Start-Up Grant (R-155-000-203-133). Wang Miao is supported by the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking Uni-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

versity) and Beijing Natural Science Foundation (Z190001). We thank the referee and Dr. Eric Tchetgen Tchetgen for helpful comments, and Drs. Zhiqiang Tan and Heng Shu for help with the data application.

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.

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.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arellano, M. and Meghir, C. (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *The Review of Economic Studies*,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

59(3):537–559.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore.

Bonet, B. (2001). Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 48–55, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bowden, R. J. and Turkington, D. A. (1990). *Instrumental Variables*, volume 8. Cambridge University Press.

Buchinsky, M., Li, F., and Liao, Z. (2018). Estimation and inference of semiparametric models using data from several sources. Technical report, Working paper.

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2003). Understanding what instrumental variables estimate: Estimating marginal and average returns to education. *processed, Univer-*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

sity of Chicago, The American Bar Foundation and Stanford University, July, 19.

Chen, X., Hong, H., Tarozzi, A., et al. (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

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.

Clarke, P. S. and Windmeijer, F. (2012). Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652.

Cui, Y. and Tchetgen Tchetgen, E. (2019). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *arXiv preprint arXiv:1911.09260*.

Currie, J. and Yelowitz, A. (2000). Are public housing projects good for kids? *Journal of public economics*, 75(1):99–124.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31.

Didelez, V., Meng, S., Sheehan, N. A., et al. (2010). Assumptions of iv methods for observational epidemiology. *Statistical Science*, 25(1):22–40.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.

Graham, B. S., Pinto, C. C. d. X., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34(2):288–301.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

in making program evaluations. *Journal of Human Resources*, pages 441–462.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kennedy, E. H., Lorch, S., and Small, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143.

Klevmarken, A. (1982). *Missing Variables and Two-stage Least-squares Estimation from More Than One Data Set*. Booklet from IUI. Industriens utredningsinstitut.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Lawlor, D. A. (2016). Commentary: Two-sample mendelian randomization: opportunities and challenges. *International Journal of Epidemiology*, 45(3):908.

Miao, W. and Tchetgen Tchetgen, E. (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *American Journal of Epidemiology*, 185(10):950–953.

Molina, J., Rotnitzky, A., Sued, M., and Robins, J. (2017). Multiple robustness in factorized likelihood models. *Biometrika*, 104(3):561–581.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

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):373–396.

Pacini, D. (2019). The two-sample linear regression model with interval-censored covariates. *Journal of Applied Econometrics*, 34(1):66–81.

Pacini, D. and Windmeijer, F. (2016). Robust inference for the two-sample 2sls estimator. *Economics Letters*, 146:50–54.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Pierce, B. L. and Burgess, S. (2013). Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. *Handbook of Econometrics*, 6:5469–5547.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, 23(8):2379–2412.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16(1):21–37.

Robins, J. M. and Greenland, S. (1996). Identification of causal effects using instrumental variables: Comment. *Journal of the American Statistical Association*, 91(434):456–458.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319.

Robins, J. M. and Rotnitzky, A. (2001). Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11:920–936.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2018). Multiply robust causal inference with double negative control adjustment for unmeasured confounding. *arXiv preprint arXiv:1808.04906*.

Shu, H. and Tan, Z. (2019). Improved methods for moment restriction models with data combination and an application to two-sample instrumental variable estimation. *Canadian Journal of Statistics*.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., and Tchetgen, E. T. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28:1965–1983.

Sun, B. and Tchetgen Tchetgen, E. J. (2018). On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association*, 113(521):369–379.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947.

Tan, Z. (2006a). A distributional approach for causal inference using propensity scores. *Journal*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

of the American Statistical Association, 101(476):1619–1637.

Tan, Z. (2006b). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618.

Tan, Z. (2007). Comment: Understanding or, ps and dr. *Statistical Science*, 22(4):560–568.

Tan, Z. (2010a). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association*, 105(489):157–169.

Tan, Z. (2010b). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *Canadian Journal of Statistics*, 38(4):609–632.

Tchetgen Tchetgen, E. J. and Robins, J. (2010). The semiparametric case-only estimator. *Biometrics*, 66(4):1138–1144.

Tchetgen Tchetgen, E. J., Robins, J. M., and Rotnitzky, A. (2009). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

Van der Laan, M. J., Laan, M., and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Wang, L., Robins, J. M., and Richardson, T. S. (2017). On falsification of the binary instru-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

mental variable model. *Biometrika*, 104(1):229–236.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT press.

Wright, P. G. (1928). *Tariff on Animal and Vegetable Oils.* Macmillan Company, New York.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2018). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Annals of Statistics*, in press.

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.

Department of Statistics and Applied Probability, National University of Singapore

E-mail: stasb@nus.edu.sg

Department of Probability and Statistics, Peking University

E-mail: mwfy@pku.edu.cn