Web Appendix to Doubly Robust Estimation of the Local Average Treatment Effect Curve

Elizabeth L. Ogburn

Johns Hopkins University, Baltimore, USA.

Andrea Rotnitzky

Di Tella University, Buenos Aires, Argentina and Harvard University, Boston, USA.

James M. Robins

Harvard University, Boston, USA.

Remark: the proofs in this Appendix are identical regardless of which $j \in \{1,2\}$ is used to defined H_j and all the related quantities, so long as the same j is used throughout. So, in what follows, the subscript j should be understood as fixed at either 1 or 2, and all the estimators and related random quantities be assumed to have been computed fixing j at the given value.

Sketch of the optimality of s_{opt} and q_{opt}

We will first prove that $\Sigma_{\mathrm{q,s}} - \Sigma_{\mathrm{q,s_{opt,j}}}$ is positive semidefinite. Define

$$A_j(\beta,\alpha) = q(V;\beta)(-1)^{1-Z} p(Z|X;\alpha)^{-1} H_j(\beta)$$

and

$$M(\beta,\alpha;s) = q(V;\beta)\, s(X)(-1)^{1-Z} p(Z|X;\alpha)^{-1}.$$

Then, $\widehat{\beta}_{\mathrm{s}}$ solves $E_n\{A_j(\beta,\widehat{\alpha}) - M(\beta,\widehat{\alpha};s)\} = 0$. Assume throughout that model (7) of the paper holds and let α_0 be the true value of α. Assume also that and $0 < \sigma_1 < p(Z=1|X) < \sigma_2 < 1$ for some σ_1 and σ_2. This condition implies that $var\{A_j(\beta,\alpha_0)\}$ and $var\{a_j(X)\}$ are both finite when, as we will assume throughout this Appendix, $var(Y), var\{q(V;\beta)\}$ and $var(m_j(V;\beta))$ are finite. It follows from standard Taylor expansion arguments for $M-$estimators that

$$\sqrt{n}\left(\widehat{\beta}_{\mathrm{s}} - \beta_0\right) = I_j^{-1}\sqrt{n}E_n\{A_{0,j} - M_0(s) - \Psi_j S\} + o_p(1)$$

where $M_0(s) = M(\beta_0,\alpha_0;s)$, $A_{0,j} = A_j(\beta_0,\alpha_0)$, $I_j = \frac{\partial}{\partial\beta'} E\{A_j(\beta,\alpha_0)\}|_{\beta=\beta_0}$, $S = \partial \log p(Z|X;\alpha)/\partial\alpha|_{\alpha=\alpha_0}$ and $\Psi_j = E[\{A_{0,j} - M_0(s)\}S']\, var(S)^{-1}$.

When $0 < \sigma_1 < p(Z=1|X) < \sigma_2 < 1$, then $A_{0,j} - M_0(s) - \Psi_j S$ has finite variance, provided, as we assume throughout that Y and $m_j(V;\beta_0)$ have finite variance. Then, by the Central Limit Theorem, $\sqrt{n}\left(\widehat{\beta}_{\mathrm{s}} - \beta_0\right)$ converges in law to a mean zero normal distribution with variance equal to

$$\Sigma_{\mathrm{q,s}} = I_j^{-1} var\left\{A_{0,j} - M_0(s) - \Psi_j S\right\}\left(I_j^T\right)^{-1}$$

Now, the space

$$\Lambda = \left\{M_0(\widetilde{s}) : \widetilde{s} \text{ arbitrary real valued}\right\}$$

is the same as the space

$$\left\{g(Z,X) : E\left\{g(Z,X)|X\right\} = 0,\ g \text{ arbitrary real valued}\right\}$$

and a quick check shows that

$$M_0(s_{\mathrm{opt,j}}) = E(A_{0,j}|X,Z) - E(A_{0,j}|X).$$

So, $M_0(s_{\mathrm{opt,j}})$ is the vector whose l^{th} element is the projection of the l^{th} entry of $A_{0,j}$ into the space Λ, i.e. $M_0(s_{\mathrm{opt,j}})$ satisfies $E\left\{\left(A_{0,j} - M_0(s_{\mathrm{opt,j}})\right)M_0(\widetilde{s})\right\} = 0$ for all $M_0(\widetilde{s}) \in \Lambda$. Then,
$\Psi_{\mathrm{opt},j} = E\left[\left\{A_{0,j} - M_0(s_{\mathrm{opt,j}})\right\}S'\right] var(S)^{-1} = 0$ because each entry of S is an element of Λ. Consequently,

$$\Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}} = I_j^{-1} var\left\{A_{0,j} - M_0(s_{\mathrm{opt,j}})\right\}\left(I_j^T\right)^{-1}.$$

Finally, $\Sigma_{\mathrm{q,s}} - \Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}} = var\left\{A_{0,j} - M_0(s) - \Psi_j S\right\} - var\left\{A_{0,j} - M_0(s_{\mathrm{opt,j}})\right\}$ ≥ 0 because each entry of the vector $M_0(s) + \Psi_j S$ is an element of Λ.

We turn now to the proof that $\Sigma_{\mathrm{q,s}_{\mathrm{opt,j}}} - \Sigma_{\mathrm{q}_{\mathrm{opt,j}},\mathrm{s}_{\mathrm{opt,j}}}$ is positive semidefinite.

For any given $q(V;\beta)$, let $q_0(V) = q(V;\beta_0)$. Also, define $\varepsilon_j(\beta) = (-1)^{1-Z} p(Z|X)^{-1}\left\{H_j(\beta) - s_{\mathrm{opt,j}}(X)\right\}$, $\varepsilon_j = \varepsilon_j(\beta_0)$, $\Delta_j(V) = E\left\{\left.\frac{\partial}{\partial\beta}\varepsilon_j(\beta)\right|_{\beta_0}\middle| V\right\}$ and $\sigma_j^2(V) = E\left(\varepsilon_j^2|V\right)$. With these definitions $A_{0,j} - M_0(s_{\mathrm{opt,j}}) = q_0(V)$

ε_j and $I_j = E\left\{\Delta_j(V)\, q_0(V)^T\right\}$. Then,

$$\begin{aligned}
\left\{\Sigma_{\mathrm{q,s_{opt,j}}}\right\}^{-1} &= E\left\{\Delta_j(V)\, q_0(V)^T\right\} E\left\{\sigma_j^2(V)\, q_0(V)\, q_0(V)^T\right\}^{-1} E\left\{q_0(V)\,\Delta_j(V)^T\right\} \\
&= E\left[\frac{\Delta_j(V)}{\sigma_j(V)}\left\{q_0(V)\,\sigma_j(V)\right\}^T\right] E\left\{\sigma_j^2(V)\, q_0(V)\, q_0(V)^T\right\}^{-1} \\
&\quad \times E\left[\left\{q_0(V)\,\sigma_j(V)\right\}\left\{\frac{\Delta_j(V)}{\sigma_j(V)}\right\}^T\right] \\
&= var\left\{\Pi\left[\Delta_j(V)/\sigma_j(V)\right|\left\langle q_0(V)\,\sigma_j(V)\right\rangle]\right\} \\
&\leq E\left[\left\{\Delta_j(V)/\sigma_j(V)\right\}\left\{\Delta_j(V)/\sigma_j(V)\right\}^T\right] \\
&= E\left\{\Delta_j(V)\, q_j^*(V)^T\right\} E\left\{\sigma_j^2(V)\, q_j^*(V)\, q_j^*(V)^T\right\}^{-1} E\left\{q_j^*(V)\,\Delta(V)^T\right\} \\
&= \left\{\Sigma_{q^*,\mathrm{s_{opt,j}}}\right\}^{-1}
\end{aligned}$$

where $q_j^*(V) = \Delta_j(V)/\sigma_j^2(V)$ and

$$\begin{aligned}
&\Pi\left[\left.\frac{\Delta_j(V)}{\sigma_j(V)}\right|\left\langle q_0(V)\,\sigma_j(V)\right\rangle\right] \\
=&E\left[\left\{\frac{\Delta_j(V)}{\sigma_j(V)}\right\}\left\{q_0(V)\,\sigma_j(V)\right\}^T\right] \times E\left\{\sigma_j^2(V)\, q_0(V)\, q_0(V)^T\right\}^{-1} q_0(V)\,\sigma_j(V)
\end{aligned}$$

is the predicted value from the population multivariate least squares regression of $\Delta_j(V)/\sigma_j(V)$ on the linear span of $q_0(V)\,\sigma_j(V)$. We therefore conclude that q_j^* is the optimal function q. To finalize the proof we need to confirm that $q^*(V) = q_{opt}(V)$.

First note that $\partial H_j(\beta)/\partial\beta = -\left\{\partial m_j(V;\beta)/\partial\beta\right\} m_j(V;\beta)^{2(1-j)}\, DY^{j-1}$. Then

$$\begin{aligned}
\Delta_j(V) &= E\left\{\left.\frac{\partial}{\partial\beta}\varepsilon_j(\beta)\right|_{\beta_0}\middle| V\right\} \\
&= E\left\{(-1)^{1-Z}\, p(Z|X)^{-1}\left.\frac{\partial}{\partial\beta}H_j(\beta)\right|_{\beta_0}\middle| V\right\} \\
&= E\left[(-1)^{1-Z}\, p(Z|X)^{-1}\left\{-\left\{\partial m_j(V;\beta)/\partial\beta\right\}|_{\beta_0}\right\} m_j(V;\beta)^{2(1-j)}\, DY^{j-1}\middle| V\right] \\
&= -\left.\frac{\partial}{\partial\beta}m_j(V;\beta)\right|_{\beta_0} m_j(V;\beta)^{2(1-j)}\, E\left\{(-1)^{1-Z}\, p(Z|X)^{-1}\, DY^{j-1}\middle| V\right\}.
\end{aligned}$$

and

$$\begin{aligned}\sigma_j(V)^2 &= E\left(\left[(-1)^{1-Z} p(Z|X)^{-1}\{H_j(\beta) - s_{\text{opt,j}}(X)\}\right]^2 \middle| V\right) \\ &= E\left(p(Z|X)^{-2}\{H_j(\beta) - s_{\text{opt,j}}(X)\}^2 \middle| V\right)\end{aligned}$$

This concludes the proof.

Sketch of the proof that $\widehat{\beta}_{\text{dr}}$ is locally efficient

We now sketch the proof that when the specifications (11) and (13) if $V \neq X$ or (14) if $V = X$ hold, and specification (7) of the paper hold, and $0 < \sigma_1 < p(Z=1|X) < \sigma_2 < 1$ for some σ_1 and σ_2, $\sqrt{n}\left(\widehat{\beta}_{\text{dr}} - \beta_0\right)$ converges in law to a mean zero normal distribution with variance equal to $\Sigma_{\text{q,s}_{\text{opt,j}}}$ where $\widehat{\beta}_{\text{dr}}$ is the doubly robust estimator defined in section 3.2 of the paper. Let $\eta^* = \text{plim}\widehat{\eta}\left(\widehat{\beta}_{\text{dr}}\right)$, $\gamma^* = \text{plim}\widehat{\gamma}$ and $s^*(X) = a(X;\alpha_0,\eta^*,\gamma^*)$. It follows from standard Taylor expansion arguments for M-estimators that under (7) and $0 < \sigma_1 < p(Z=1|X) < \sigma_2 < 1$ for some σ_1 and σ_2,

$$\sqrt{n}\left(\widehat{\beta}_{\text{dr}} - \beta_0\right) = I_j^{-1}\sqrt{n}E_n\left\{A_{0,j} - M_0(s^*) - \Psi_j^* S\right\} + o_p(1)$$

where $M_0(s), A_{0,j}, I_j$ and S are defined as in the preceding proof and $\Psi_j^* = E\left[\{A_{0,j} - M_0(s^*)\}S'\right] var(S)^{-1}$. When, in addition, (11) and (13) hold if $V \neq X$ or (14) if $V = X$ hold we also have $a(X;\alpha_0,\eta^*,\gamma^*) = s_{\text{opt,j}}(X)$, from where the claim follows because as showed in the preceding proof, $var\left\{A_{0,j} - M_0(s_{\text{opt,j}}) - \Psi_j^* S\right\} = \Sigma_{\text{q,s}_{\text{opt,j}}}$.

Sketch of the proof that $\widetilde{\beta}_{\text{dr}}$ is doubly-robust and has the efficiency property stated in section 3.4

Assume throughout that $0 < \sigma_1 < p(Z=1|X) < \sigma_2 < 1$ for some σ_1 and σ_2. With $A_j(\beta,\alpha)$ and $M(\beta,\alpha;s)$ defined as in the first proof the estimator $\widetilde{\beta}_{\text{dr}}$ solves $E_n[A_j(\beta,\widehat{\alpha})] - \widehat{C}(\beta)^{\text{T}} E_n[M(\beta,\widehat{\alpha};\widehat{a}(\beta))] = 0$ where $\widehat{a}(\beta) \equiv a(X;\widehat{\alpha},\widehat{\eta}(\beta),\widehat{\gamma})$. Let $\overline{\beta}_{\text{dr}}$ be the solution of $E_n[A_j(\beta,\alpha)] - C^*(\beta)^{\text{T}} E_n[M(\beta,\widehat{\alpha};\widehat{a}(\beta))] = 0$ where $C^*(\beta) = \text{plim}\widehat{C}(\beta)$. Under regularity conditions, $n^{1/2}\left(\widetilde{\beta}_{\text{dr}} - \overline{\beta}_{\text{dr}}\right) = o_p(1)$, so to show the claims made on $\widetilde{\beta}_{\text{dr}}$ it suffices to show that the same claims hold for $\overline{\beta}_{\text{dr}}$.

To show that $\overline{\beta}_{\text{dr}}$ is doubly robust, first note that under the propensity score model (7) given in the paper, $E_n[M(\beta,\widehat{\alpha};\widehat{a}(\beta))] = o_p(1)$, so $\overline{\beta}_{\text{dr}}$ converges to the solution of $E[A_j(\beta,\alpha_0)] = 0$ which is precisely β_0. On the other hand, to show the consistency of $\overline{\beta}_{\text{dr}}$ when the specifications

given in displays (11) and (13) if $V \neq X$ or (14) if $V = X$ hold, it suffices to show that in such case $C^*(\beta_0)$ is the identity matrix (since then $\overline{\beta}_{\mathrm{dr}}$ is equal to the double robust estimator $\widehat{\beta}_{\mathrm{dr}}$ described in section 3.2 of the paper. To show that $C^*(\beta_0)$ is the identity matrix note that it is the $p \times p$ matrix formed by the first p columns of the $p \times (p+r)$ matrix $D_1 D_2^{-1}$ where

$$D_1 = E\left\{A_j(\beta_0, \alpha^*) K(\beta_0)\right\} \tag{1}$$

and

$$D_2 = E\left(\left[\begin{array}{c} q(V;\beta_0)(-1)^{1-Z} p(Z|X;\alpha^*)^{-1} h(Z,X;\eta_0) \\ \partial \log p(Z|X;\alpha)/\partial\alpha|_{\alpha=\alpha^*} \end{array}\right] \times K(\beta_0)\right) \tag{2}$$

and $\alpha^* = \mathrm{plim}\widehat{\alpha}$. When the specifications given in displays (11) and (13) if $V \neq X$ or (14) if $V = X$ hold, $\mathrm{plim}\widehat{\eta}(\beta_0) = \eta_0$ and $\mathrm{plim}\widehat{\gamma} = \gamma_0$ where (η_0, γ_0) satisfy $h(Z,X;\eta_0;\gamma_0) = E(H_j|Z,X)$. Then,

$$D_1 = E\left\{q(V;\beta_0)(-1)^{1-Z} p(Z|X;\alpha^*)^{-1} h(Z,X;\eta_0) K(\beta_0)\right\}$$

Thus $C^*(\beta_0)$ is the $p \times p$ identity matrix because D_1 is the $p \times (p+r)$ upper block of the $(p+r) \times (p+r)$ matrix D_2.

We will now prove that when the propensity score parametric specification (7) of the paper holds then $\widetilde{\beta}_{\mathrm{dr}}$ is at least as efficient asymptotically as any estimator $\widehat{\beta}_C$ solving $E_n[A_j(\beta,\widehat{\alpha})] - C^{\mathrm{T}} E_n[M(\beta,\widehat{\alpha};\widehat{a}(\beta))] = 0$ for an arbitrary $p \times p$ matrix C. Henceforth assume model (7) holds and let α_0 be the true value of α. Let $\eta^*(\beta) = \mathrm{plim}\widehat{\eta}(\beta)$, $\gamma^* = \mathrm{plim}\widehat{\gamma}$, $a^*(x;\beta) = \mathrm{plim} a(x;\widehat{\alpha},\eta^*(\beta),\gamma^*)$ and $M_0^* = q(V;\beta_0) a_j^*(X;\beta_0)(-1)^{1-Z} p(Z|X;\alpha_0)^{-1}$. It follows from standard Taylor expansion arguments for $M-$estimators that

$$\sqrt{n}\left(\widehat{\beta}_C - \beta_0\right) = I_j^{-1}\sqrt{n} E_n\left\{A_{0,j} - C M_0^* - \Psi_{C,j} S\right\} + o_p(1)$$

where $A_{0,j}, I_j$ and S are defined as in the first proof of this Appendix and $\Psi_{C,j} = E\left[\left\{A_{0,j} - C M_0^*\right\} S'\right] var(S)^{-1}$. On the other hand,

$$\sqrt{n}\left(\overline{\beta}_{\mathrm{dr}} - \beta_0\right) = I_j^{-1}\sqrt{n} E_n\left\{A_{0,j} - C^* M_0^* - \Psi_{C^*,j} S\right\} + o_p(1)$$

where $C^* = C^*(\beta_0)$, is the $p \times p$ matrix formed by the first p columns of the $p \times (p+r)$ matrix $D_1 D_2^{-1}$ with D_1 and D_2 as in display (2) with α_0 instead of α^*. So, it suffices to show that

$$var\left\{A_{0,j} - C M_0^* - \Psi_{C,j} S\right\} - var\left\{A_{0,j} - C^* M_0^* - \Psi_{C^*,j} S\right\} \geq 0 \tag{3}$$

When model (7) holds, $D_1 D_2^{-1}$ is the same as $B = E\left\{A_{0,j}\left[M_0^{*\mathrm{T}}, S^{\mathrm{T}}\right]\right\} var\left\{\left[M_0^{*\mathrm{T}}, S^{\mathrm{T}}\right]^{\mathrm{T}}\right\}^{-1}$, i.e. the least squares constant in the population regression of A_{0j} on M_0^* and S. Furthermore, write

$B = [B_1, B_2]$ where B_1 is $p \times p$ and B_2 is $p \times r$. Then, by definition, $C^* = B_1$ and consequently, by the minimum variance property of the regression residual, we have

$$var\left\{A_{0,j} - C^* M_0^* - B_2 S\right\} \leq var\left\{A_{0,j} - C^* M_0^* - \Psi_{C^*,j} S\right\}$$

On the other hand, by definition, $\Psi_{C^*,j}$ is the least squares constant in the population regression of $A_{0,j} - C^* M_0$ on S. Therefore

$$var\left\{A_{0,j} - C^* M_0^* - \Psi_{C^*,j} S\right\} \leq var\left\{A_{0,j} - C^* M_0^* - B_2 S\right\}$$

We thus conclude that $var\left\{A_{0,j} - C^* M_0^* - \Psi_{C^*,j} S\right\} = var\left\{A_{0,j} - B_1 M_0^* - B_2 S\right\}$. Now, by definition of $B = [B_1, B_2]$, $var\left\{A_{0,j} - B_1 M_0^* - B_2 S\right\} - var\left\{A_{0,j} - B_1^* M_0^* - B_2^* S\right\}$ is negative semidefinite for any conformable matrices B_1^* and B_2^*, in particular, for the choices $B_1^* = C$ and $B_2^* = \Psi_{C,j}$. This proves that (3) is positive semidefinite.

Best least squares approximations when the parametric specifications for LATE(V) and MLATE(V) are incorrect

Let $G_1(\beta) \equiv E\left[w(V)\left\{LATE(V) - m_1(V;\beta)\right\}^2 | D_1 > D_0\right]$ and $\beta_{w,0} \equiv \arg\min_\beta G_1(\beta)$. Assuming $G_1(\beta)$ is differentiable at $\beta_{w,0}$ and that differentiation can be exchanged with integration we have that $\beta_{w,0}$ solves $Q_1(\beta) = 0$ where $Q_1(\beta) = E\left[\frac{\partial m_1(V;\beta)}{\partial \beta} w(V)\left\{LATE(V) - m_1(V;\beta)\right\}\middle| D_1 > D_0\right]$. Now, under the IV assumptions we have

$$\begin{aligned}
& Q_1(\beta) P(D_1 > D_0) \\
= \; & E\left[\frac{\partial m_1(V;\beta)}{\partial \beta} w(V)\left\{LATE(V) - m_1(V;\beta)\right\}\middle| D_1 > D_0\right] P(D_1 > D_0) \\
= \; & E\left[\frac{\partial m(V;\beta)}{\partial \beta} w(V)\left\{LATE(V) - m_1(V;\beta)\right\} E(D_1 - D_0|X)\right] \\
= \; & E\left[\frac{(-1)^{1-Z}}{p(Z|X)} \frac{\partial m(V;\beta)}{\partial \beta} w(V)\left\{IV(V) - m_1(V;\beta)\right\} D\right] \\
= \; & E\left[\frac{(-1)^{1-Z}}{p(Z|X)} \frac{\partial m(V;\beta)}{\partial \beta} w(V)\left\{H_1(\beta) - H_1\right\}\right] \\
= \; & E\left\{\frac{(-1)^{1-Z}}{p(Z|X)} \frac{\partial m(V;\beta)}{\partial \beta} w(V) H_1(\beta)\right\}
\end{aligned}$$

thus showing that $\beta_{w,0}$ satisfies equation (22) of the paper.

An estimator $\widehat{\widehat{\beta}}_{\text{opt,dr}}$ which satisfies property (a) of section 3.4 of the paper and which has limiting normal distribution with variance equal to

$\Sigma_{\mathrm{q_{opt,1}},\mathrm{s_{opt,1}}}$ when conditions (i) and (ii) of section 3.3 of the paper hold and the specifications (19) and (20) of the paper are correct, and yet it converges to a weighted least squares approximation when the parametric specification for $LATE\left(V\right)$ is wrong is computed just as $\widehat{\widehat{\beta}}_{\mathrm{dr}}$ in section 3.3 of the paper but replacing in every step of its construction $q_w\left(v;\beta\right)$ with the function $\widehat{q}_{\mathrm{opt},1}\left(v;\beta\right)$ defined in section 3.3. When the specification for $LATE\left(V\right)$ is wrong $\widehat{\widehat{\beta}}_{\mathrm{opt,dr}}$ converges in probability to $\beta_{w^*,0}$ where $w^*\left(V\right)=e_1\left(V;\delta^*\right)t_1\left(V;\omega^*\right)^{-1}$ with $\delta^*=\operatorname{plim}\widehat{\delta}$ and $\omega^*=\operatorname{plim}\widehat{\omega}$.

Next, let $G_2\left(\beta\right)\equiv E\left[e_0\left(V\right)w\left(V\right)\left\{<LATE\left(V\right)-m_2\left(V;\beta\right)\right\}^2|D_1>D_0\right]$ where $e_0\left(V\right)=E\left(Y_0|D_1>D_0,V\right)$ and redefine $\beta_{w,0}\equiv\arg\min_\beta G_2\left(\beta\right)$. Assuming $G_2\left(\beta\right)$ is differentiable at $\beta_{w,0}$ and that differentiation can be exchanged with integration we have that $\beta_{w,0}$ solves $Q_2\left(\beta\right)=0$ where $Q_2\left(\beta\right)=E\left[\frac{\partial m_2(V;\beta)}{\partial\beta}e_0\left(V\right)w\left(V\right)\left\{MLATE\left(V\right)-m_2\left(V;\beta\right)\right\}\middle|D_1>D_0\right]$. Under the IV assumptions

$$
\begin{aligned}
&Q_2\left(\beta\right)P\left(D_1>D_0\right)\\
=&E\left[w\left(V\right)\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}E\left(Y_0|D_1>D_0,V\right)\left\{MIV\left(V\right)-m_2\left(V;\beta\right)\right\}\middle|D_1>D_0\right]P\left(D_1>D_0\right)\\
=&E\left[w\left(V\right)\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}MIV\left(V\right)^{-1}E\left(Y_1|D_1>D_0,V\right)\left\{MIV\left(V\right)-m_2\left(V;\beta\right)\right\}\middle|D_1>D_0\right]\\
&\times P\left(D_1>D_0\right)\\
=&E\left[w\left(V\right)\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}MIV\left(V\right)^{-1}E\left\{Y_1\times\left(D_1-D_0\right)|V\right\}\left\{MIV\left(V\right)-m_2\left(V;\beta\right)\right\}\right]\\
=&E\left[w\left(V\right)\frac{\left(-1\right)^{1-Z}}{p\left(Z|X\right)}\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}MIV\left(V\right)^{-1}YD\left\{MIV\left(V\right)-m_2\left(V;\beta\right)\right\}\right]\\
=&E\left[w\left(V\right)\frac{\left(-1\right)^{1-Z}}{p\left(Z|X\right)}\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}m_2\left(V;\beta\right)Y\left\{m_2\left(V;\beta\right)^{-D}-MIV\left(V\right)^{-D}\right\}\right]\\
=&E\left[w\left(V\right)\frac{\left(-1\right)^{1-Z}}{p\left(Z|X\right)}\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}m_2\left(V;\beta\right)\left\{H_2\left(\beta\right)-H_2\right\}\right]\\
=&E\left\{w\left(V\right)\frac{\left(-1\right)^{1-Z}}{p\left(Z|X\right)}\frac{\partial m_2\left(V;\beta\right)}{\partial\beta}m_2\left(V;\beta\right)H_2\left(\beta\right)\right\}
\end{aligned}
$$

The form of the quantity inside the last expectation agrees with the form of the quantity inside the last expectation of the previous display, except that $w\left(V\right)$ is replaced with $w\left(V\right)m_2\left(V;\beta\right)$ and the subscript 1 is replaced with the subscript 2. Thus, the estimator $\widehat{\widehat{\beta}}_{\mathrm{dr}}$ of section 3.4 computed using $H_2\left(\beta\right)$ instead of $H_1\left(\beta\right)$ and with $q_w\left(V;\beta\right)$ redefined as $m_2\left(V;\beta\right)\times\left\{\partial m_2\left(V;\beta\right)/\partial\beta\right\}\times w\left(V\right)$ satisfies the properties claimed in the paper.

Proof of the restrictions imposed by the IV assumptions (our model) and by the Robins-Tan model.

First we prove that the only restrictions on the observed data law imposed by assumptions (i)-(vi) (i.e. our model, as defined in section 4 of the paper) beyond $0 < P(Z = 1|X) < 1$ are

$$\Pr(y < Y \leq y', D = 1|, Z = 1, X) - \Pr(y < Y \leq y', D = 1|, Z = 0, X) \geq 0, \tag{4}$$

$$\Pr(y < Y \leq y', D = 0|, Z = 0, X) - \Pr(y < Y \leq y', D = 0|, Z = 1, X) \geq 0 \tag{5}$$

and

$$E\{E(D|Z = 1, X)|V\} - E\{E(D|Z = 0, X)|V\} > 0. \tag{6}$$

The proof that the inequalities (4), (5) and (6) are implied by assumptions (i)-(vi) hinges on the following identity (Imbens and Angrist, 1994), which holds for any $g(\cdot)$ under the IV assumptions (i), (ii), (v) and (vi)

$$\begin{aligned} &E\{g(Y, D, X)|Z = 1, X\} - E\{g(Y, D, X)|Z = 0, X\} \\ &\quad = E[C\{g(Y_1, D_1, X) - g(Y_0, D_0, X)\}|X] \end{aligned} \tag{7}$$

where $C = D_1 - D_0$. Letting $g(Y, D, X) = D$ gives that (6) is equivalent to $E(C|V) > 0$, which follows from instrumentation (iv) and monotonicity (v). Letting $g(Y, D, X) = I(y \leq Y \leq y', D = 1)$ gives that (4) is the same as $\Pr(y \leq Y_1 \leq y', C = 1|X) \geq 0$, which holds because probabilities are non-negative. Likewise, letting $g(Y, D, X) = -I(y \leq Y \leq y', D = 0)$ gives that (5) is also the same as $\Pr(y \leq Y_0 \leq y', C = 1|X) \geq 0$.

To show that (4), (5) and (6) and $0 < P(Z = 1|X) < 1$ are the only restrictions imposed on the observed data law by assumptions (i) - (vi), first note that assumptions (i)-(v) determine a model, denoted herein as $\mathcal{A}_0$, on the law of $W = (Y_0, Y_1, D_0, D_1, Z, X)$, defined by the restrictions

A.1 $(Y_0, Y_1, D_0, D_1) \amalg Z|X$,
A.2 $E(D_1|V) - E(D_0|V) > 0$ with probability 1,
A.3 $D_1 - D_0 \geq 0$,
A.4 $0 < P(Z = 1|X) < 1$.

Under assumption (vi), the received treatment D is equal to the function $u(D_0, D_1, Z)$ of (D_0, D_1, Z) defined by

$$u(D_0, D_1, Z) \equiv \begin{cases} 1 \text{ if } D_0 = 1 \text{ or if } (Z = 1 \text{ and } D_1 - D_0 = 1) \\ 0 \text{ if } D_1 = 0 \text{ of if } (Z = 0 \text{ and } D_1 - D_0 = 1) \end{cases}$$

and the observed outcome Y is equal to the function $s(W)$ of W defined by

$$s(W) = \begin{cases} Y_1 \text{ if } u(D_0, D_1, Z) = 1 \\ Y_0 \text{ if } u(D_0, D_1, Z) = 0. \end{cases}$$

So, under (vi) we have that the joint distribution of (Y, D) given (Z, X) is determined by the equalities

$$\begin{aligned} &\Pr(y < Y \leq y', D = d | Z = z, X) \qquad (8) \\ &= \begin{cases} \Pr(y < Y_1 \leq y', D_1 = 1 | Z = 1, X) \text{ if } (d, z) = (1, 1) \\ \Pr(y < Y_1 \leq y', D_0 = 1 | Z = 0, X) \text{ if } (d, z) = (1, 0) \\ \Pr(y < Y_0 \leq y', D_1 = 0 | Z = 1, X) \text{ if } (d, z) = (0, 1) \\ \Pr(y < Y_0 \leq y', D_0 = 0 | Z = 0, X) \text{ if } (d, z) = (0, 0) \end{cases} \end{aligned}$$

Thus, to study the restrictions imposed on the conditional distribution of (Y, D) given (Z, X) by (i)-(vi) we examine the restrictions imposed by (A.1)-(A.4) on the conditional probabilities in the set

$$\{\Pr(y < Y_d \leq y', D_z = d | Z = z, X) : y \text{ and } y' \in \mathcal{Y}\ ,\ z \text{ and } d \in \{0, 1\}\} \qquad (9)$$

Assumption A.1 imposes the sole restriction that

$$\begin{aligned} &\Pr(y < Y_d \leq y', D_z = d | Z = z, X) \qquad (10) \\ = \ &\Pr(y < Y_d \leq y', D_z = d | X), \text{ for } y \text{ and } y' \in \mathcal{Y}\ ,\ z \text{ and } d \in \{0, 1\}, \end{aligned}$$

So the set (9) is the same as the set

$$\{\Pr(y < Y_d \leq y', D_z = d | X) : y \text{ and } y' \in \mathcal{Y}\ ,\ z \text{ and } d \in \{0, 1\}\} \qquad (11)$$

On this set, assumption A.3 imposes the sole restrictions

$$\Pr(y < Y_1 \leq y', D_1 = 1 | X) \geq \Pr(y < Y_1 \leq y', D_0 = 1 | X) \text{ for } y \text{ and } y' \in \mathcal{Y} \qquad (12)$$

$$\Pr(y < Y_0 \leq y', D_1 = 0 | X) \leq \Pr(y < Y_0 \leq y', D_0 = 0 | X) \text{ for } y \text{ and } y' \in \mathcal{Y} \qquad (13)$$

Inequality (12) and inequality (10) with $d = 1$ imply the sole restriction (4), whereas inequality (13) and inequality (10) with $d = 0$ imply the sole restriction (5). Finally, taking $y = -\infty$ and $y' = +\infty$ in (8), assumption A.2 implies the sole restriction (6).

Next we prove that the only restriction on the observed data law imposed by assumptions (i)-(iv), (v-ATT), and (vi) (i.e. by the Robins-Tan model as defined in section 4 of the paper) is

$$E\{P(D = 1 | Z = 1, X) | V\} \neq E\{P(D = 1 | Z = 0, X) | V\} \text{ with probability 1.} \qquad (14)$$

Assumptions (i)-(iv) and (v-ATT) impose a model $\mathcal{B}_0$ on the law of $W = (Y_0, Y_1, D_0, D_1, Z, X)$ defined by the restrictions

B.1 $(Y_0, Y_1, D_0, D_1) \amalg Z|X$,
B.2 $E(D_1|V) - E(D_0|V) \neq 0$ with probability 1,
B.3 $0 < P(Z = 1|X) < 1$.

Just as deduced earlier, assumption (vi) implies that the conditional distribution of (Y, D) given (Z, X) is related to the counterfactual data law through (8). So, assumption B.1 once again implies that the set (9) and (11) are equal. On this set, assumption B.2 imposes only a restriction on the choice $y = -\infty$ and $y' = +\infty$, namely

$$E[\Pr(D_1 = 1|X)|V] \neq E[\Pr(D_0 = 1|X)|V]$$

Taking $y = -\infty$ and $y' = +\infty$ and $d = 1$ in (8) we conclude that the only restriction implied by simultaneously assuming B.1 and B.2 on the observed data law is (14).

Proof of point (c) of section 4 of the paper

By point (b) of section 4 of the paper, the intersection model imposes the same restrictions on the observed data law as our model. According to the results in the preceding proof, these restrictions are the inequality constraints (4), (5) and (6). On the other hand, also according to the results in the preceding proof, the Robins-Tan model imposes solely the inequality constraint (14). This proofs the assertion. Note that (14) is implied by constraint (6) so, as far as models for the observed data law, our model is indeed a submodel of the Robins-Tan model.

Proof of the variation independence of $E[\underline{\varphi}(X)|V], E[H_1|Z, X]$ and $p(Z|X)$ with $IV(V)$ and of $E[\underline{\varphi}(X)|V], E[H_2|Z, X], p(Z|X)$ with $MIV(V)$

Let Y be real or integer valued and have unbounded support. Let F_O be a given observed data law satisfying $0 < P(Z = 1|X) < 1$ and the restrictions in the preceding displays (4), (5) and (6) (i.e. those implied by the IV assumptions (i)-(vi)). Suppose that $IV(v)$ is a given function of v, say $m_1(v)$. Then, by definition, $H_1 = Y - m_1(V)D$ and by construction, H_1 satisfies the restriction

$$E\{[E(H_1|Z = 1, X) - E(H_1|Z = 0, X)|V]\} = 0. \tag{15}$$

This restriction is the same regardless of what the function $m_1(\cdot)$ is. Also, restrictions (5), (4) and (6) of the preceding subsection are equivalent to the restrictions that for all $y < y'$ in the real line,

$$\begin{aligned}&\Pr(y - m_1(V) < H_1 \leq y' - m_1(V), D = 1|Z = 1, X)\\ &- \Pr(y - m_1(V) < H_1 \leq y' - m_1(V), D = 1|Z = 0, X) \geq 0\end{aligned}$$

and

$$\Pr\left(y < H_1 \leq y', D = 0 | Z = 0, X\right) - \Pr\left(y < H_1 \leq y', D = 0 | Z = 1, X\right) \geq 0$$

which is equivalent to the restriction that for all $h < h'$ in the real line,

$$\Pr\left(h < H_1 \leq h', D = 1 | Z = 1, X\right) - \Pr\left(h < H_1 \leq h', D = 1 | Z = 0, X\right) \geq 0$$

$$\Pr\left(h < H_1 \leq h', D = 0 | Z = 0, X\right) - \Pr\left(h < H_1 \leq h', D = 0 | Z = 1, X\right) \geq 0$$

This restriction is the same regardless of what the function $m_1(\cdot)$ is. Finally, regardless of what $m_1(\cdot)$ is, H_1 has unbounded support because Y has unbounded support. So, all restrictions on the law of H_1 are the same regardless of what the value of $m_1(\cdot)$ is. Consequently, the range of possible values taken by $E(H_1|Z, X)$ is the same regardless of what $m_1(\cdot)$ is. This, in turn, implies that the set of permissible laws of $X|V$ is the same regardless of the functional form $m_1(\cdot)$ as (15) is the only restriction on the law of $X|V$. Finally, the functional $IV(V)$ depends only on the law of $(Y, D)|Z, X$ and on the law of $X|V$ but not on the law of $p(Z|X)$, which proofs then that $IV(V)$ is variation independent with $p(Z|X)$.

Next consider Y with support equal to either $[0, \infty)$ or the non-negative integers. Let F_O be a given observed data law satisfying the restrictions implied by assumptions (i)-(vii). These restrictions are $0 < P(Z = 1|X) < 1$, the restrictions (5), (4) and (6) of the preceding subsection (i.e. those implied by the IV assumptions (i)-(vi)) and the restriction

$$E\left[Y(1 - D)|Z = 1, X\right] - E\left[Y(1 - D)|Z = 1, X\right] \neq 0 \tag{16}$$

implied by (vii) (which follows by taking in (7) $g(Y, D, X) = Y(1 - D)$).

Suppose that $MIV(v)$ is a given function of v, say $m_2(v)$. Then, $H_2 = Ym_2(V)^{-D}$ and by construction, H_2 satisfies the restriction

$$E\left\{\left[E(H_2|Z = 1, X) - E(H_2|Z = 0, X)|V\right]\right\} = 0$$

This restriction is the same regardless of what the function $m_2(\cdot)$ is. By definition, $m_2(v) > 0$ because Y has support equal or included in $[0, \infty)$. Thus, restrictions (5), (4) and (6) of the preceding subsection are equivalent to the restrictions that for all non-negative reals $y < y'$,

$$\Pr\left(ym_2(V)^{-1} < H_2 \leq y'm_2(V)^{-1}, D = 1 | Z = 1, X\right)$$
$$- \Pr\left(ym_1(V)^{-1} < H_2 \leq y'm_2(V)^{-1}, D = 1 | Z = 0, X\right) \geq 0$$

and

$$\Pr\left(y < H_2 \leq y', D = 0 | Z = 0, X\right) - \Pr\left(h < H_2 \leq h', D = 0 | Z = 1, X\right) \geq 0$$

which, by virtue of $m_2(V)$ being positive, are equivalent to the restrictions that for all non-negative reals $h < h'$,

$$\Pr(h < H_2 \leq h', D = 1|, Z = 1, X) - \Pr(h < H_2 \leq h', D = 1|, Z = 0, X) \geq 0$$

$$\Pr(h < H_2 \leq h', D = 0|Z = 0, X) - \Pr(h < H_2 \leq h', D = 0|Z = 1, X) \geq 0$$

These restrictions are the same regardless of what the function $m_2(\cdot)$ is. Likewise, restriction (16) is equivalent to restriction

$$E[H_2(1-D)|Z=1,X] - E[H_2(1-D)|Z=1,X] \neq 0 \qquad (17)$$

because $H_2(1-D) = Y(1-D)$. This restriction is not affected by the value of $m_2(\cdot)$.

Finally, H_2 has support equal or included in $[0, \infty)$ because Y does so and by definition $m_2(v) > 0$. So, all restrictions on H_2 are the same regardless of what the value of $m_2(\cdot)$. Consequently, the range of possible values taken by $E(H_2|Z,X)$ is the same regardless of what $m_2(\cdot)$. The proof of the variation independence of $f(X|V)$ and $p(Z|X)$ with $MIV(V)$ is identical to the one given above for the variation independence of $f(X|V)$ and $p(Z|X)$ with $IV(V)$.

Structural interpretations of $E(H_j|z,X)$ under our model and under the Robins-Tan model.

We will prove that under the Robins-Tan model,

$$E(H_1|z,X) = E(Y_0|X) - \{ATT(V) - ATT(z,X)\} \Pr(D_z = 1|X).$$

and under our model,

$$\begin{aligned} & E(H_1|z,X) \\ = \; & E(Y_0|X) + \{E(Y_0 - Y_1|X, T = at) - LATE(X)\} \Pr(T = at|X) \\ & + \{LATE(X) - LATE(V)\} \{z \Pr(T \in \{at,\ co\}|X) + (1-z) \Pr(T = ne|X)\}. \end{aligned}$$

Under the Robins-Tan model, we know that

$$E\{Y - ATT(Z,X) D|Z,X\} = E(Y_0|Z,X)$$

Adding and subtracting $ATT(V)D$ and using the fact that $E(Y_0|Z,X) = E(Y_0|X)$ we have

$$E[H_1 + \{ATT(V) - ATT(Z,X)\} D|Z,X] = E[Y_0|X]$$

where $H_1 = Y - ATT(V)D$, or equivalently

$$\begin{aligned} E(H_1|Z,X) &= E(Y_0|X) - \{ATT(V) - ATT(Z,X)\} E(D|Z,X) \\ &= E(Y_0|X) - \{ATT(V) - ATT(Z,X)\} E(D_Z|Z,X) \end{aligned}$$

So,

$$\begin{aligned} E\left(H_1|Z=z,X\right) &= E\left(Y_0|X\right)-\left\{ATT\left(V\right)-ATT\left(z,X\right)\right\}E\left(D_z|Z=z,X\right) \\ &= E\left(Y_0|X\right)-\left\{ATT\left(V\right)-ATT\left(z,X\right)\right\}E\left(D_z|X\right) \end{aligned}$$

thus showing that

$$E\left(H_1|z,X\right)=E\left(Y_0|X\right)-\left\{ATT\left(V\right)-ATT\left(z,X\right)\right\}\Pr\left(D_z=1|X\right).$$

Now we show that under our model

$$\begin{aligned} & E\left(H_1|z,X\right) \\ = \; & E\left(Y_0|X\right)+\left\{E\left(Y_0-Y_1|X,T=at\right)-LATE\left(X\right)\right\}\Pr\left(T=at|X\right) \\ & +\left\{LATE\left(X\right)-LATE\left(V\right)\right\}\left\{z\Pr\left(T\in\left\{at,\ co\right\}|X\right)+\left(1-z\right)\Pr\left(T=ne|X\right)\right\}. \end{aligned}$$

Let T be the variable denoting compliance type. We have

$$\begin{aligned} & E\left\{Y-LATE\left(X\right)D|Z,X\right\}= \\ = \; & E\left\{Y-LATE\left(X\right)D|T=co,Z,X\right\}\Pr\left(T=co|Z,X\right)+ \\ & E\left\{Y-LATE\left(X\right)D|T=at,Z,X\right\}\Pr\left(T=at|Z,X\right)+ \\ & E\left\{Y-LATE\left(X\right)D|T=ne,Z,X\right\}\Pr\left(T=ne|Z,X\right) \\ = \; & E\left\{Y-LATE\left(X\right)D|T=co,Z,X\right\}\Pr\left(T=co|Z,X\right)+ \\ & \left\{E\left(Y_1|T=at,X\right)-LATE\left(X\right)\right\}\Pr\left(T=at|X\right)+ \\ & E\left(Y_0|T=ne,X\right)P\left(T=ne|X\right) \\ = \; & E\left(Y_0|T=co,X\right)P\left(T=co|X\right)+ \\ & \left\{E\left(Y_1|T=at,X\right)-LATE\left(X\right)\right\}P\left(T=at|X\right)+ \\ & E\left\{Y_0|T=ne,X\right\}P\left(T=ne|X\right) \\ = \; & E\left(Y_0|X\right)-E\left(Y_0|T=at,X\right)P\left(T=at|X\right) \\ & +\left\{E\left(Y_1|T=at,X\right)-LATE\left(X\right)\right\}P\left(T=at|X\right) \\ = \; & E\left(Y_0|X\right)+\left\{E\left(Y_1-Y_0|T=at,X\right)-LATE\left(X\right)\right\}P\left(T=at|X\right). \end{aligned}$$

Next, note that

$$\begin{aligned} & E\left\{Y-LATE\left(V\right)D|Z,X\right\}= \\ = \; & E\left\{Y-LATE\left(X\right)D|Z,X\right\}+E\left[\left\{LATE\left(X\right)-LATE\left(V\right)\right\}D|Z,X\right] \\ = \; & E\left\{Y-LATE\left(X\right)D|Z,X\right\}+\left\{LATE\left(X\right)-LATE\left(V\right)\right\}E\left(D|Z,X\right). \end{aligned}$$

Furthermore, $E\left(D|Z=1,X\right)=P\left(T=at\text{ or}T=\ co|X\right)$ and $E\left(D|Z=0,X\right)=P\left(T=ne|X\right)$. Thus, with $H_1=Y-D\times LATE\left(V\right)$

$$\begin{aligned} & E\left(H_1|z,X\right) \\ = \; & E\left\{Y-LATE\left(v\right)D|Z=z,X\right\} \\ = \; & E\left(Y_0|X\right)+\left\{E\left(Y_0-Y_1|X,T=at\right)-LATE\left(X\right)\right\}\Pr\left(T=at|X\right) \\ & +\left\{LATE\left(X\right)-LATE\left(V\right)\right\}\left\{z\Pr\left(T\in\left\{at,\ co\right\}|X\right)+\left(1-z\right)\Pr\left(T=ne|X\right)\right\}. \end{aligned}$$

References

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