Biometrika (2017), *pp*. 1–9

Printed in Great Britain

Online Supplement to "Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder"

BY WANG MIAO

Guanghua School of Management, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, P.R.C.

mwfy@pku.edu.cn

ZHI GENG

School of Mathematical Sciences, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, P.R.C.

zhigeng@pku.edu.cn

AND ERIC TCHETGEN TCHETGEN

Department of Biostatistics, Harvard University, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

etchetge@hsph.harvard.edu

This supplement includes a counterexample illustrating lack of identification of the error mechanism $\text{pr}(w \mid u)$ for model (f); a counterexample showing non-identification of $\text{pr}\{y \mid \text{do}(x)\}$ for model (a); proof and extension of Theorems 1–2; simulations evaluating the performance of the testing strategy; discussion on existence of a solution to the integral equation (6) and proof of Proposition 1; and computation details for Example 1.

A.1. NON-IDENTIFICATION OF $\text{pr}(w \mid u)$ IN MODEL (f)

Suppose the true data generating mechanism $\text{pr}(x, y, u, z, w)$ is encoded by the following probability matrices: for $i = 1, 2$,

$$P(X) = \begin{pmatrix} 5 \\ 5 \end{pmatrix} /10, \quad P(Z \mid X) = \begin{pmatrix} 14 & 9 \\ 7 & 12 \end{pmatrix} /21, \quad P(U \mid Z, x_i) = \begin{pmatrix} 8 - 2i & 9 - i \\ 2 + 2i & 1 + i \end{pmatrix} /10,$$

$$P(y_1, W \mid U, x_i) = \begin{pmatrix} 6i & 22i - 14 \\ 24i + 16 & 4i + 20 \end{pmatrix} /100, \; P(y_2, W \mid U, x_i) = \begin{pmatrix} 20 - 6i & 54 - 22i \\ 64 - 24i & 40 - 4i \end{pmatrix} /100$$

then we have

$$P(W \mid U, x_i) = P(W \mid U) = \begin{pmatrix} 2 & 4 \\ 8 & 6 \end{pmatrix} /10.$$

Thus, the data generating process satisfies model (f). The true causal effect is $\text{pr}\{y_1 \mid \text{do}(x_i)\} = 0.122 + 0.285i$ for $i = 1, 2$. The distribution of observed variables is captured by $P(X), P(Z \mid X)$ and $P(y, W \mid Z, x) = P(y, W \mid U, x)P(U \mid Z, x)$ for $y = y_1, y_2$ and $x = x_1, x_2$.

However, we can construct a different data generating process with identical observed data distribution. Letting

$$A = \begin{pmatrix} 1.1 & -0.4 \\ -0.1 & 1.4 \end{pmatrix},$$

and

$$P_2(U \mid Z, x_i) = A^{-1}P(U \mid Z, x_i) = \begin{pmatrix} 24-4i & 26-2i \\ 6+4i & 4+2i \end{pmatrix}/30,$$

$$P_2(y_1, W \mid U, x_i) = P(y_1, W \mid U, x_i)A = \begin{pmatrix} 7+22i & 142i-98 \\ 78+130i & 108-20i \end{pmatrix}/500,$$

$$P_2(y_2, W \mid U, x_i) = P(y_1, W \mid U, x_i)A = \begin{pmatrix} 83-22i & 338-142i \\ 332-130i & 152+20i \end{pmatrix}/500,$$

then the new data generating process $\text{pr}_2(x, y, u, w, z)$ encoded by $\{P_2(U \mid Z, x_i), P_2(y, W \mid U, x_i), P_2(X) = P(X), P_2(Z \mid X) = P(Z \mid X)\}$ satisfies model (f) with

$$P_2(W \mid U, x_i) = P_2(W \mid U) = P(W \mid U)A = \begin{pmatrix} 18 & 48 \\ 82 & 52 \end{pmatrix}/100,$$

which is different from $P(W \mid U)$. However, the distribution of observed variables remains the same, because for all (x, y),

$$P_2(y, W \mid Z, x) = P_2(y, W \mid U, x)P_2(U \mid Z, x) = P(y, W \mid Z, x).$$

Therefore, $P(W \mid U)$ cannot be identified. But applying (5), the causal effect is identified by $\text{pr}\{y_1 \mid \text{do}(x_i)\} = 0.122 + 0.285i$ for $i = 1, 2$, which can also be verified from $\text{pr}(x, y, u, z, w)$.

A.2. NON-IDENTIFICATION OF $\text{pr}\{y \mid \text{do}(x)\}$ IN (a)

When condition (i) in the article is not met, for example, when only one proxy is available, the causal effect $\text{pr}\{y \mid \text{do}(x)\}$ is in general not identifiable. We illustrate with a counterexample below. Suppose the true data generating mechanism $\text{pr}(x, y, u, z)$ for (a) is determined by the following probability matrices

$$P(X) = \begin{pmatrix} 5 \\ 5 \end{pmatrix}/10, \qquad P(U \mid X) = \begin{pmatrix} 8 & 9 \\ 2 & 1 \end{pmatrix}/10, \qquad P(Z \mid U) = \begin{pmatrix} 2 & 4 \\ 8 & 6 \end{pmatrix}/10,$$

$$P(Y \mid U, x_1) = \begin{pmatrix} 2 & 6 \\ 8 & 4 \end{pmatrix}/10, \qquad P(Y \mid U, x_2) = \begin{pmatrix} 3 & 5 \\ 7 & 5 \end{pmatrix}/10.$$

The causal effect is $\text{pr}\{y_1 \mid \text{do}(x_i)\} = (7i + 19)/100$ for $i = 1, 2$. Letting

$$\Lambda(U \mid x) = \begin{pmatrix} \text{pr}(u_1 \mid x) & 0 \\ 0 & \text{pr}(u_2 \mid x) \end{pmatrix} \text{ for } x = x_1, x_2,$$

the observed variable distribution is captured by $P(X)$ and $P(Y, Z \mid x) = P(Y \mid U, x)\Lambda(U \mid x)P(Z \mid U)^{\text{T}}$ for $x = x_1, x_2$.

We construct a new data generating process $\text{pr}_2(x, y, u, z)$ with

$$P_2(U \mid X) = \begin{pmatrix} 5 & 4 \\ 5 & 6 \end{pmatrix}/10, \qquad P_2(Z \mid U) = \begin{pmatrix} 34 & 14 \\ 66 & 86 \end{pmatrix}/100,$$

$$P_2(Y \mid U, x_1) = \begin{pmatrix} 41 & 15 \\ 59 & 85 \end{pmatrix} /100, \qquad P_2(Y \mid U, x_2) = \begin{pmatrix} 37 & 29 \\ 63 & 71 \end{pmatrix} /100,$$

and $P_2(X) = P(X)$. Letting

$$\Lambda_2(U \mid x) = \begin{pmatrix} \text{pr}_2(u_1 \mid x) & 0 \\ 0 & \text{pr}_2(u_2 \mid x) \end{pmatrix} \text{ for } x = x_1, x_2,$$

we have $P_2(Y \mid U, x) = P(Y \mid U, x)A(x)$, $P_2(Z \mid U) = P(Z \mid U)B^{\mathrm{T}}$ and $\Lambda_2(U \mid x) = A(x)^{-1}\Lambda(U \mid x)B^{-1}$ for $x = x_1, x_2$, with

$$A(x_1) = \begin{pmatrix} 48 & 112 \\ 52 & -12 \end{pmatrix} /100, \quad A(x_2) = \begin{pmatrix} 675 & 1050 \\ 325 & -50 \end{pmatrix} /1000, \quad B = \begin{pmatrix} 3 & 7 \\ 13 & -3 \end{pmatrix} /10.$$

The new data generating process results in identical distribution of observed variables, because for $x = x_1, x_2$,

$$\begin{aligned} P_2(Y, Z \mid x) &= P_2(Y \mid U, x)\Lambda_2(U \mid x)P_2(Z \mid U)^{\mathrm{T}} \\ &= P(Y \mid U, x)\Lambda(U \mid x)P(Z \mid U)^{\mathrm{T}} \\ &= P(Y, Z \mid x); \end{aligned}$$

but with a different causal effect $\text{pr}_2\{y_1 \mid \text{do}(x_i)\} = (5.7i + 21)/100$ for $i = 1, 2$.

A.3. Proof of Theorem1

THEOREM 1. *Assuming model (f) and condition (ii), for any solution $h(w, x, y)$ to (6),*

$$\begin{aligned} \text{pr}(y \mid u, x) &= \int_{-\infty}^{+\infty} h(w, x, y) f(w \mid u) dw, \\ \text{pr}\{y \mid \text{do}(x)\} &= \int_{-\infty}^{+\infty} h(w, x, y) f(w) dw. \end{aligned}$$

Proof. For any (x, y), suppose $h(w, x, y)$ solves (6): for all z,

$$\text{pr}(y \mid z, x) = \int_{-\infty}^{+\infty} h(w, x, y) f(w \mid z, x) dw,$$

then under model (f), for all z,

$$\int_{-\infty}^{+\infty} \text{pr}(y \mid u, x) f(u \mid z, x) du = \int_{-\infty}^{+\infty} h(w, x, y) \left\{ \int_{-\infty}^{+\infty} f(w \mid u) f(u \mid z, x) du \right\} dw.$$

Under the completeness condition (ii), we must have

$$\text{pr}(y \mid u, x) = \int_{-\infty}^{+\infty} h(w, x, y) f(w \mid u) dw;$$

taking expectation over u on both sides, we obtain

$$\text{pr}\{y \mid \text{do}(x)\} = \int_{-\infty}^{+\infty} h(w, x, y) f(w) dw.$$

A.4. PROOF OF THEOREM 2 AND EXTENSION

To prove Theorem 2, we need the following lemma, which is Theorem 1.12 of Shao (2003).

LEMMA 1. *Let $X_1, X_2, \ldots$ and Y be random k-vectors satisfying $a_n(X_n - c) \to Y$ in distribution, where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to+\infty} a_n = +\infty$. Let g be a function from $\mathbb{R}^k$ to $\mathbb{R}$. Suppose that g has continuous partial derivatives of order $m > 1$ in a neighborhood of c, with all the partial derivatives of order smaller than $m - 1$ vanishing at c, but with the mth-order partial derivatives not all vanishing at c. Then*

$$a_n^m\{g(X_n) - g(c)\} \to \frac{1}{m!}\sum_{i_1=1}^{k}\cdots\sum_{i_m=1}^{k}\left.\frac{\partial^m g}{\partial x_{i_1}\cdots\partial x_{i_m}}\right|_{x=c} Y_{i_1}\times\cdots\times Y_{i_m} \text{ in distribution,}$$

where Y_j is the jth component of Y.

Lemma 1 concerns the approximate distribution for a function of a series of variable/vector that converge in distribution. We use Slutsky's theorem and Lemma 1 to prove Theorem 2.

THEOREM 2. *Assuming model (f), conditions (iv) and (10)–(11), if $\mathbb{H}_0$ is correct, then $n^{1/2}\xi_y \to N(0, \Omega_y)$ in distribution, with $\Omega_y = I - \Sigma_y^{-1/2}Q^{\mathrm{T}}(Q\Sigma_y^{-1}Q^{\mathrm{T}})^{-1}Q\Sigma_y^{-1/2}$ of rank $r = ij - k$, and $T_y \to \chi_r^2$ in distribution.*

Proof of Theorem 2. 1. Given that $\widehat{Q} \to Q$, $\widehat{\Sigma}_y \to \Sigma$ in probability and $n^{1/2}(\widehat{q}_y - q_y) \to N(0, \Sigma_y)$ in distribution, applying Slutsky's theorem, we have $n^{1/2}(\xi_y - \Omega_y\Sigma_y^{-1/2}q_y) \to N(0, \Omega_y)$ in distribution with $\Omega_y = I - \Sigma_y^{-1/2}Q^{\mathrm{T}}(Q\Sigma_y^{-1}Q^{\mathrm{T}})^{-1}Q\Sigma_y^{-1/2}$. If $\mathbb{H}_0$ is correct, then $q_y^{\mathrm{T}} = P(y \mid U)P(W \mid U)^{-1}Q$, and thus $\Omega_y\Sigma_y^{-1/2}q_y = 0$, which implies that $n^{1/2}\xi_y \to N(0, \Omega_y)$ in distribution. Because $Q\Sigma_y^{-1/2}$ has rank k, $\Sigma_y^{-1/2}Q^{\mathrm{T}}(Q\Sigma_y^{-1}Q^{\mathrm{T}})^{-1}Q\Sigma_y^{-1/2}$ is an idempotent matrix (Banerjee & Roy, 2014, Corollary 11.5) of rank k, i.e., it has k eigenvalues equal to one and $ij - k$ eigenvalues equal to zero. Hence, Ω_y is an idempotent matrix of rank $r = ij - k$.

2. For fixed y, applying Lemma 1 with $g(x) = x^{\mathrm{T}}x$, we have

$$T_y = g(n^{1/2}\xi_y) \to N(0, \Omega_y)^{\mathrm{T}}N(0, \Omega_y) \text{ in distribution.}$$

Because Ω_y is an idempotent matrix of rank $r = ij - k$, there exists a unitary matrix V such that $V\Omega_y V^{\mathrm{T}} = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$, a diagonal matrix with r eigenvalues equal to one. Thus, $VN(0, \Omega_y) \sim N\{0, \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)\}$, and

$$N(0, \Omega_y)^{\mathrm{T}}N(0, \Omega_y) = \{VN(0, \Omega_y)\}^{\mathrm{T}}\{VN(0, \Omega_y)\} \sim \chi_r^2.$$

Therefore, $T_y \to \chi_r^2$ in distribution.

Theorem 2 can be generalized to account for all levels of a categorical Y. Consider an l-category outcome and let $q^{\mathrm{T}} = (q_1^{\mathrm{T}}, \ldots, q_{l-1}^{\mathrm{T}})$; then under $\mathbb{H}_0$, we have

$$q = \{P(y_1 \mid U)P(W \mid U)^{-1}, \ldots, P(y_{l-1} \mid U)P(W \mid U)^{-1}\}\begin{pmatrix} Q & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & Q \end{pmatrix}.$$

In the below, we use Q_0 to denote the diagonal matrix on the right hand side, which is a $k(l-1) \times ij(l-1)$ matrix and has full row rank. A test statistic aggregating all levels of Y

can be constructed by replacing $(\widehat{q}_y, \widehat{Q})$ with $(\widehat{q}, \widehat{Q}_0)$ whenever they appear in the construction of ξ_y and T_y.

Suppose we have estimators $(\widehat{q}, \widehat{Q}_0)$ that satisfy

$$n^{1/2}(\widehat{q} - q) \to N(0, \Sigma) \text{ in distribution,} \tag{12}$$

$$\widehat{Q}_0 \to Q_0 \text{ and } \widehat{\Sigma} \to \Sigma \text{ in probability, with } \widehat{\Sigma}, \Sigma \text{ positive-definite.} \tag{13}$$

We let

$$\xi = \{I - \widehat{\Sigma}^{-1/2}\widehat{Q}_0^{\mathrm{T}}(\widehat{Q}_0\widehat{\Sigma}^{-1}\widehat{Q}_0^{\mathrm{T}})^{-1}\widehat{Q}_0\widehat{\Sigma}^{-1/2}\}\widehat{\Sigma}^{-1/2}\widehat{q},$$

and propose the test statistic $T = n\xi^{\mathrm{T}}\xi$.

Following the proof of Theorem 2, we have the corollary.

COROLLARY 1. *Assuming model (f), conditions (iv) and* (12)–(13)*, if* $\mathbb{H}_0$ *is correct, then* $n^{1/2}\xi \to N(0, \Omega)$ *in distribution, with* $\Omega = I - \Sigma^{-1/2}Q_0^{\mathrm{T}}(Q_0\Sigma^{-1}Q_0^{\mathrm{T}})^{-1}Q_0\Sigma^{-1/2}$ *of rank* $r(l-1)$*, and* $T \to \chi^2_{r(l-1)}$ *in distribution.*

Aggregating all levels of an l-category outcome leads to a chi-square test with $r(l-1)$ degrees of freedom. For a continuous Y, discretization is required to perform the proposed chi-square test, however, in many situations where the average causal effect is of interest, one can use $q = \{E(Y \mid Z, x_1), \ldots, E(Y \mid Z, x_i)\}^{\mathrm{T}}$ in construction of the test statistic and perform the test on the mean scale.

A.5. SIMULATIONS FOR THE TESTING STRATEGY

We conduct simulations to evaluate the performance of the proposed testing strategy. We consider two data generating mechanisms that satisfy model (c). In the first case, we set

$$P(X) = \begin{pmatrix} 3 \\ 3 \\ 4 \end{pmatrix}/10, \qquad P(U \mid X) = \begin{pmatrix} 3 & 6 & 5 \\ 7 & 4 & 5 \end{pmatrix}/10, \qquad P(W \mid U) = \begin{pmatrix} 8 & 3 \\ 2 & 7 \end{pmatrix}/10,$$

$$P(Y \mid U, x_1) = \begin{pmatrix} 5 & 4 \\ 3 & 2 \\ 2 & 4 \end{pmatrix}/10, \quad P(Y \mid U, x_2) = \begin{pmatrix} 4 & 6 \\ 2 & 3 \\ 4 & 1 \end{pmatrix}/10, \quad P(Y \mid U, x_3) = \begin{pmatrix} 3 & 2 \\ 4 & 5 \\ 3 & 3 \end{pmatrix}/10.$$

The causal effect is nonzero but cannot be identified.

In the second case, $\{P(X), P(U \mid X), P(W \mid U)\}$ remain the same as in the first case, but $P(Y \mid U, x)$ does not vary with x,

$$P(Y \mid U, x_1) = P(Y \mid U, x_2) = P(Y \mid U, x_3) = \begin{pmatrix} 5 & 4 \\ 3 & 5 \\ 2 & 1 \end{pmatrix}/10.$$

Hence, $\mathbb{H}_0$ holds in this case.

Under each case, we generate 1000 datasets under sample sizes $200, 400$ and 600. In construction of the test statistics $T_{y=1}$ and T, we use empirical probability mass functions to estimate $\mathrm{pr}(w \mid z, x)$ and $\mathrm{pr}(y \mid z, x)$. Under the settings we consider, $i = 3, j = 1, k = 2, l = 3$ (Z treated as a constant), and thus, $T_{y=1}$ and T lead to a χ^2_1 and a χ^2_2 test, respectively.

Table 1 presents the power of the tests when using $T_{y=1}$ and T as the test statistic, respectively. The significance level is 0.05. When $\mathbb{H}_0$ does not hold, the tests have good empirical power that

increases toward unity as the sample size increases; when $\mathbb{H}_0$ holds, the empirical type I error is close to the nominal level of 0.05. Under the settings we consider, $T_{y=1}$ does not result in a substantial loss of power compared to T. Such results confirm that the proposed tests perform reasonably well when the sample size is moderate.

Table 1: Power of the test

	T		$T_{y=1}$	
Sample size	$\mathbb{H}_0$ correct	$\mathbb{H}_0$ incorrect	$\mathbb{H}_0$ correct	$\mathbb{H}_0$ incorrect
200	0.052	0.748	0.048	0.695
400	0.057	0.952	0.046	0.943
600	0.062	0.994	0.055	0.994

A.6. DISCUSSION ON EXISTENCE OF A SOLUTION TO (6)

Equation (6) is a Fredholm integral equation of the first kind. A conventional and rigorous approach to study this problem is the singular value decomposition (Kress, 1989, Theorem 15.16) developed in functional analysis. The approach has previously been used by statisticians and economists (Carrasco et al., 2007; Darolles et al., 2011) to study nonparametric instrumental regression. In the following, we introduce the singular value decomposition of a compact operator and Picard's theorem to describe an if and only if condition for existence of a solution, and then apply Picard's theorem to prove Proposition 1 in the article.

According to Kress (1989, Theorem 15.16), given Hilbert spaces H_1 and H_2, a compact operator $K: H_1 \longmapsto H_2$ and its adjoint operator $K^*: H_2 \longmapsto H_1$, there exists a singular system $(\lambda_n, \varphi_n, \psi_n)_{n=1}^{+\infty}$ of K with nonzero singular values $\{\lambda_n\}$ and orthogonal sequences $\{\varphi_n \in H_1\}$ and $\{\psi_n \in H_2\}$ such that

$$K\varphi_n = \lambda_n \psi_n, \quad K^*\psi_n = \lambda_n \varphi_n.$$

By the means of singular value decomposition, the following result known as Picard's theorem (Kress, 1989, Theorem 15.18) characterizes if and only if conditions for existence of a solution to the corresponding Fredholm integral equation of the first kind.

LEMMA 2 (PICARD'S THEOREM). *Letting $K: H_1 \longmapsto H_2$ be a compact operator with singular system $(\lambda_n, \varphi_n, \psi_n)_{n=1}^{+\infty}$, given $\phi \in H_2$, the equation of the first kind $Kh = \phi$ is solvable if and only if*

1. $\phi \in \mathcal{N}(K^*)^{\perp}$; *and*
2. $\sum_{n=1}^{+\infty} \lambda_n^{-2} |\langle \phi, \psi_n \rangle|^2 < +\infty$;

where $\mathcal{N}(K^*) = \{h : K^*h = 0\}$ is the null space of K^*, and $^{\perp}$ denotes the orthogonal complement to a subset.

We apply Lemma 2 to prove Proposition 1 in the article. We let $L^2\{F(t)\}$ denote the space of all square integrable functions of t with respect to a cumulative distribution function $F(t)$, which is a Hilbert space with the inner product

$$\langle g, h \rangle = \int_{-\infty}^{+\infty} g(t)h(t)dF(t) \text{ for all } g, h \in L^2\{F(t)\}.$$

Letting

$$K(w,z,x) = \frac{f(w,z \mid x)}{f(w \mid x)f(z \mid x)},$$

for any x, we define the linear operators

$$\begin{aligned}
&K_x :\ L^2\{F(w \mid x)\} \longmapsto L^2\{F(z \mid x)\},\\
&K_x h = \int_{-\infty}^{+\infty} K(w,z,x)h(w)dF(w \mid x) = E\{h(w) \mid z,x\},\ h \in L^2\{F(w \mid x)\},\\
&K_x^* :\ L^2\{F(z \mid x)\} \longmapsto L^2\{F(w \mid x)\},\\
&K_x^* g = \int_{-\infty}^{+\infty} K(w,z,x)g(z)dF(z \mid x) = E\{g(z) \mid w,x\},\ g \in L^2\{F(z \mid x)\},
\end{aligned}$$

which are integral operators with kernel $K(w,z,x)$ and are referred to as conditional expectation operators (Carrasco et al., 2007, Example 2.3, page 5656). One can verify that K_x^* is in fact the adjoint operator of K_x by checking

$$\langle K_x h, g\rangle = \langle h, K_x^* g\rangle = E\{h(w)g(z) \mid x\}.$$

Proposition 1 in the article is an immediate corollary of Lemma 2 by noting that under the regularity condition (v), K_x is a compact operator, and under the completeness condition (iii), $\mathcal{N}(K_x^*)^\perp = L^2\{F(z \mid x)\}$.

Proof of Proposition 1. First, we note that K_x is a compact operator by assuming $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(w \mid z,x)f(z \mid w,x)dwdz < +\infty$ (Carrasco et al., 2007, Example 2.3, page 5659). Thus, there exists a singular value decomposition $(\lambda_{x,n}, \varphi_{x,n}, \psi_{x,n})_{n=1}^{+\infty}$ of K_x according to Kress (1989, Theorem 15.16) and Carrasco et al. (2007, Theorem 2.41). Second, under the completeness condition (iii), we prove $\mathcal{N}(K_x^*)^\perp = L^2\{F(z \mid x)\}$ by showing $\mathcal{N}(K_x^*) = \{g = 0\}$. For any $g \in \mathcal{N}(K_x^*)$, we have $K_x^* g = E\{g(z) \mid w,x\} = 0$ almost surely; from condition (iii), we must have $g(z) = 0$ almost surely. As a result, $\mathcal{N}(K_x^*) = \{g = 0\}$, and therefore, $\mathcal{N}(K_x^*)^\perp = L^2\{F(z \mid x)\}$. Third, assuming $\int_{-\infty}^{+\infty} \text{pr}^2(y \mid z,x)f(z \mid x)dz < +\infty$ for any given (x,y), we must have $\text{pr}(y \mid z,x) \in L^2\{F(z \mid x)\}$, and thus $\text{pr}(y \mid z,x) \in \mathcal{N}(K_x^*)^\perp$. Last, together with (vii), Lemma 2 implies existence of a solution to (6). □

A.7. Computation details for example 1

Under the setting of Example 1 in the article, for all (x,y), we solve $h(w,x,y)$ from the integral equation:

$$\text{pr}(y \mid z,x) = \int_{-\infty}^{+\infty} f(w \mid z,x)h(w,x,y)dw,$$

with

$$f(w \mid z,x) = \frac{1}{\sigma(x)}\phi\left\{\frac{w - \beta_0(x) - \beta_1(x)z}{\sigma(x)}\right\}.$$

By substitution $z' = \{\beta_0(x) + \beta_1(x)z\}/\sigma(x)$, $w' = w/\sigma(x)$, and by letting

$$g(y,z',x) = \text{pr}\left\{y \mid z = \frac{z'\sigma(x) - \beta_0(x)}{\beta_1(x)}, x\right\},$$

we can solve h from

$$g(y,z',x)=\int_{-\infty}^{+\infty}\phi(z'-w')h\{w'\sigma(x),x,y\}dw',$$

which is an integral equation of convolution type, and can be solved by applying the Fourier transform. Letting h_1 and h_2 denote the Fourier transforms of ϕ and g respectively:

$$\begin{aligned}h_1(v)&=\int_{-\infty}^{+\infty}\exp(-\mathrm{i}vz')\phi(z')dz'\\&=\int_{-\infty}^{+\infty}\exp(-\mathrm{i}vz)\phi(z)dz,\\h_2(v,x,y)&=\int_{-\infty}^{+\infty}\exp(-\mathrm{i}vz')g(y,z',x)dz'\\&=\frac{\beta_1(x)}{\sigma(x)}\int_{-\infty}^{+\infty}\exp\left\{-\mathrm{i}v\frac{\beta_0(x)+\beta_1(x)z}{\sigma(x)}\right\}\mathrm{pr}(y\mid z,x)dz,\end{aligned}$$

with $\mathrm{i}=(-1)^{1/2}$ the imaginary unity, we have

$$h_2(v,x,y)=h_1(v)\times\int_{-\infty}^{+\infty}\exp(-\mathrm{i}vw')h\{w'\sigma(x),x,y\}dw',$$

$$\int_{-\infty}^{+\infty}\exp(-\mathrm{i}vw')h\{w'\sigma(x),x,y\}dw'=\frac{h_2(v,x,y)}{h_1(v)};$$

by Fourier inversion, we have

$$h\{w'\sigma(x),x,y\}=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp(\mathrm{i}vw')\frac{h_2(v,x,y)}{h_1(v)}dv;$$

by substitution $w=w'\sigma(x)$, we obtain

$$h(w,x,y)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp\left\{\frac{\mathrm{i}vw}{\sigma(x)}\right\}\frac{h_2(v,x,y)}{h_1(v)}dv.$$

When $f(x,y,u,w,z)$ follows a joint normal model, one first applies linear regression to the observed variables to obtain $f(y\mid z,x)\sim N(\alpha_0+\alpha_1z+\alpha_2x,\sigma_1^2)$, $f(w\mid z,x)\sim N(\beta_0+\beta_1z+\beta_2x,\sigma_2^2)$ and $f(w)\sim N(\mu,\sigma_3^2)$; then one can verify that (6) has a unique solution

$$h(w,x,y)=\frac{1}{\sigma_4}\phi\left(\frac{y-\gamma_{01}-\alpha_1/\beta_1w-\gamma_1x}{\sigma_4}\right),$$

with $\gamma_1=\alpha_2-\alpha_1\beta_2/\beta_1$, $\gamma_{01}=\alpha_0-\alpha_1\beta_0/\beta_1$ and $\sigma_4^2=\sigma_1^2-\alpha_1^2\sigma_2^2/\beta_1^2$. The casual effect is

$$\mathrm{pr}\{y\mid\mathrm{do}(x)\}=\int_{-\infty}^{+\infty}h(w,x,y)f(w)dw=\frac{1}{\sigma}\phi\left(\frac{y-\gamma_0-\gamma_1x}{\sigma}\right),$$

with $\gamma_0=\gamma_{01}+\alpha_1\mu/\beta_1$ and $\sigma^2=\sigma_4^2+\alpha_1^2\sigma_3^2/\beta_1^2$.

In linear structural models, the path coefficient $\partial E(y\mid u,x)/\partial x$ is of interest. From $f(y\mid u,x)=\int_{-\infty}^{+\infty}h(w,x,y)f(w\mid u)dw$, the path coefficient is identified by:

$$\frac{\partial E(y\mid u,x)}{\partial x}=\int_{-\infty}^{+\infty}\frac{\partial\int_{-\infty}^{+\infty}yh(w,x,y)dy}{\partial x}f(w\mid u)dw=\gamma_1,$$

which is in fact consistent with the result of Kuroki & Pearl (2014) obtained via variance analysis

$$\frac{\partial E(y \mid u, x)}{\partial x} = \frac{\sigma_{yz}\sigma_{xw} - \sigma_{xy}\sigma_{wz}}{\sigma_{xz}\sigma_{xw} - \sigma_{xx}\sigma_{wz}},$$

where σ_{xy} denotes the covariance of X and Y, and similar notation for other variables. One can verify this by noting

$$\sigma_{yz} = \alpha_1\sigma_{zz} + \alpha_2\sigma_{xz}, \qquad \sigma_{yx} = \alpha_1\sigma_{xz} + \alpha_2\sigma_{xx},$$
$$\sigma_{wz} = \beta_1\sigma_{zz} + \beta_2\sigma_{xz}, \qquad \sigma_{wx} = \beta_1\sigma_{xz} + \beta_2\sigma_{xx}.$$

REFERENCES

BANERJEE, S. & ROY, A. (2014). *Linear Algebra and Matrix Analysis for Statistics*. Boca Raton: Taylor & Francis.

CARRASCO, M., FLORENS, J. P. & RENAULT, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, J. J. Heckman & E. Leamer, eds., vol. 6B. Amsterdam: Elsevier, pp. 5633–5751.

DAROLLES, S., FAN, Y., FLORENS, J. P. & RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–1565.

KRESS, R. (1989). *Linear Integral Equations*. Berlin: Springer.

KUROKI, M. & PEARL, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101**, 423–437.

SHAO, J. (2003). *Mathematical Statistics*. New York: Springer, 2nd ed.