Supplementary Material

Testing for causal effect for binary data when propensity scores are estimated through Bayesian Networks

Paola Vicard[1], Paola Maria Vittoria Rancoita[2], Federica Cugnata[2], Alberto Briganti[2], Fulvia Mecatti[3], Clelia Di Serio[2], Pier Luigi Conti[4*]

[1]Roma Tre University, Rome, Italy.
[2]Vita-Salute San Raffaele University, Milan, Italy.
[3]University of Milano-Bicocca, Milan, Italy.
[4]Sapienza University of Rome, Rome, Italy.

A Estimation of Weighted Average Treatment Effect

Let g be a real-valued function, playing the role of weight function, and define

$$\begin{aligned}\zeta_k &= \frac{1}{E[g(\boldsymbol{X})]} E\left[\frac{Y I_{(T=k)}}{p_k(\boldsymbol{X})} g(\boldsymbol{X})\right] \\ &= E\left[\frac{Y I_{(T=k)}}{p_k(\boldsymbol{X})} g(\boldsymbol{X})\right] \Bigg/ \sum_{\boldsymbol{x}} g(\boldsymbol{x}) P(\boldsymbol{X}=\boldsymbol{x}),\ k=0,\, 1. \end{aligned} \tag{1}$$

The Weighted Average Treatment Effect (WATE) is defined as

$$\Delta_{WATE} = \zeta_1 - \zeta_0. \tag{2}$$

Due to unconfoundedness, the following relationship hold:

$$\Delta_{WATE} = \frac{1}{E[g(\boldsymbol{X})]} E\left[E[g(\boldsymbol{X}) E[Y_{(1)} - Y_{(0)}|\boldsymbol{X}]\right]. \tag{3}$$

To estimate ζ_k, a reasoning similar to that of Section 3 of the paper can be used. Since ζ_k is the solution of the equation

$$E\left[g(\boldsymbol{X})\left(\frac{Y I_{(T=k)}}{p_k(\boldsymbol{X})} - \zeta_k\right)\right] = 0$$

it is natural to estimate ζ_k through the estimator

$$\widehat{\zeta}_k = \sum_{i=1}^{n} g(\boldsymbol{X}_i) \frac{Y_i I_{(T_i=k)}}{p_k(\boldsymbol{X}_i)}, \ k = 0, 1. \tag{4}$$

In the sequel, the vector notation

$$\boldsymbol{\zeta} = \begin{bmatrix} \zeta_0 \\ \zeta_1 \end{bmatrix}, \ \widehat{\boldsymbol{\zeta}} = \begin{bmatrix} \widehat{\zeta}_0 \\ \widehat{\zeta}_1 \end{bmatrix}.$$

will be used.

Proposition A.1. *Under the assumpions A1-A3, there exists a sequence of* i.i.d. *bivariate random vectors*

$$\boldsymbol{l}^*(Y_i, T_i, \boldsymbol{X}_i) = \begin{bmatrix} l_0^*(Y_i, T_i, \boldsymbol{X}_i) \\ l_1^*(Y_i, T_i, \boldsymbol{X}_i) \end{bmatrix}$$

with zero expectation, having the following properties.

1. $\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta})$ *and* $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{l}^*(Y_i, T_i, \boldsymbol{X}_i)$ *possess the same limiting distribution, as* $n \to \infty$.
2. *If* $E[\boldsymbol{l}^*(Y_i, T_i, \boldsymbol{X}_i)^2] > 0$, *Then* $\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) \xrightarrow{d} \boldsymbol{W}_1$ *as* $n \to \infty$

where $\boldsymbol{W}_1$ *is a bivariate r.v. with Normal* $N_2(\boldsymbol{0}, \boldsymbol{\Gamma})$ *distribution,* i.e. *with null mean vector and covariance matrix*

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix} = E\left[\boldsymbol{l}^*(Y_i, T_i, \boldsymbol{X}_i)\, \boldsymbol{l}^*(Y_i, T_i, \boldsymbol{X}_i)^T\right].$$

Proof. The proof is similar to that of Proposition 3. Define first

$$l_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i) = \frac{1}{E[g(\boldsymbol{X})]} g(\boldsymbol{X}_i) \left\{ \left(\frac{Y_i I_{(t_i=k)}}{p_k(\boldsymbol{X}_i)} - \zeta_k \right) - \frac{\theta_k(\boldsymbol{X}_i)}{p_k(\boldsymbol{X}_i)} \left(I_{(t_i=k)} - p_k(\boldsymbol{X}_i) \right) \right\},\ k = 0,\, 1.$$

Using exactly the arguments in Proposition 3, it is seen that

$$\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{l}^*(Y_i,\, T_i,\, \boldsymbol{X}_i) + o_p(1).$$

The proof now follows from the bivariate Central Limit Theorem. □

As a consequence of the above result, the asymptotic normality of WATE is obtained. Let $\boldsymbol{a}$ be the vector with components 1, -1, and define

$$\widehat{\Delta}_{WATE} = \widehat{\zeta}_1 - \widehat{\zeta}_0.$$

Proposition A.2. *Suppose that assumptions A1-A3 are fulfilled, and that* $E[\boldsymbol{l}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^2] > 0$. *Then:*

$$\sqrt{n}(\widehat{\Delta}_{WATE} - \Delta_{WATE}) \xrightarrow{d} N(0,\, \sigma_w^2) \text{ as } n \to \infty$$

where $\sigma_W^2 = \boldsymbol{a}^T \Gamma \boldsymbol{a}$.

B Estimation of Average Treatment Effect for treated

Average Treatment Effect for Treated (ATT, for short) is defined as

$$\Delta_{ATT} = E[Y_{(1)}|T = 1] - E[Y_{(0)}|T = 1]. \tag{5}$$

From unconfoundedness, it is not difficult to see that

$$\begin{aligned} E[Y_{(1)}|T=1] &= E\left[Y_{(1)} I_{(T=1)} + Y_{(0)} I_{(T=0)}|\right] = E[Y|T=1] \\ &= E\left[E\left[Y|T=1,\, \boldsymbol{X}\right]|T=1\right] \\ &= E[\theta_1(\boldsymbol{X})|T=1] = \sum_{\boldsymbol{x}} \theta_1(\boldsymbol{X}) P(\boldsymbol{X}=\boldsymbol{x}|T=1) \\ &= \frac{1}{P(T=1)} \sum_{\boldsymbol{x}} \theta_1(\boldsymbol{X}) P(\boldsymbol{X}=\boldsymbol{x},\, T=1) \end{aligned}$$

$$= \frac{1}{P(T=1)} \sum_{\boldsymbol{x}} \theta_1(\boldsymbol{X}) p_1(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}).$$

Hence, denoting by p_1 the (marginal) probability $P(T = 1)$, the relationship

$$E[Y_{(1)}|T=1] = \frac{1}{p_1} E\left[\frac{Y I_{(T=1)}}{p_1(\boldsymbol{X})} p_1(\boldsymbol{X})\right] = \frac{1}{p_1} E\left[Y I_{(T=1)}\right] \tag{6}$$

is obtained. In a similar way, and again using unconfoundedness, it is seen that

$$\begin{aligned} E[Y_{(0)}|T=1] &= \frac{1}{p_1} \sum_{\boldsymbol{x}} \theta_0(\boldsymbol{X}) p_1(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{1}{p_1} \left[\frac{Y I_{(T=0)}}{p_0(\boldsymbol{X})} p_1(\boldsymbol{X})\right]. \end{aligned} \tag{7}$$

Thus, taking into account that

$$p_1 = P(T=1) = \sum_{\boldsymbol{x}} p_1(\boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x})$$

and denoting by ξ_1, ξ_0 the quantities

$$\xi_k = E\left[\frac{Y I_{(T=k)}}{p_k(\boldsymbol{X})} p_1(\boldsymbol{X})\right] \bigg/ p_1, \ k = 0, 1 \tag{8}$$

it is seen that ATT is a special case of WATE, corresponding to a weight function $g(\boldsymbol{x}) = p_1(\boldsymbol{x})$.

If $p_1(\boldsymbol{x})$ were known, statistical inference on ξ_k could be obtained by specializing the results in Section *A*. Since it is unknown, we have to slightly change our approach. Define the estimator of (8) as solution of the equation

$$\sum_{i=1}^{n} \left(\frac{Y_i I_{(T_i=k)}}{\widehat{p}_k(\boldsymbol{X}_i)} - \xi_k \right) \widehat{p}_1(\boldsymbol{X}_i)$$

namely

$$\widehat{\xi}_k = \sum_{i=1}^{n} \frac{Y_i I_{(T_i=k)}}{\widehat{p}_k(\boldsymbol{X}_i)} \widehat{p}_1(\boldsymbol{X}_i) \bigg/ \sum_{i=1}^{n} \widehat{p}_1(\boldsymbol{X}_i), \ k = 0, 1. \tag{9}$$

Similarly to Section *A*, in the sequel the following notation will be used:

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_0 \\ \xi_1 \end{bmatrix}, \quad \widehat{\boldsymbol{\xi}} = \begin{bmatrix} \widehat{\xi}_0 \\ \widehat{\xi}_1 \end{bmatrix}.$$

Proposition B.1. *Under the assumpions A1-A3, there exists a sequence of* i.i.d. *bivariate random vectors*

$$\boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i) = \begin{bmatrix} q_0^*(Y_i,\, T_i,\, \boldsymbol{X}_i) \\ q_1^*(Y_i,\, T_i,\, \boldsymbol{X}_i) \end{bmatrix}$$

with zero expectation, having the following properties.

1. $\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi})$ *and* $\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)$ *possess the same limiting distribution, as* $n \to \infty$.
2. *If* $E[\boldsymbol{q}_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^2] > 0$ *for* $k = 0,\, 1,,$ *then* $\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}) \xrightarrow{d} \boldsymbol{W}_2$ *as* $n \to \infty$

where $\boldsymbol{W}_2$ *is a bivariate r.v. with Normal* $N_2(\mathbf{0},\, \boldsymbol{\Psi})$ *distribution,* i.e. *with null mean vector and covariance matrix*

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{bmatrix} = E\left[\boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)\, \boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^T\right].$$

Proof. The proof is similar in many respects to that of Proposition *A*.1. In the first place, the same arguments as in Proposition *A*.1 show that

$$\sqrt{n}(\widehat{\xi}_k - \xi_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{p_1} \widehat{p}_1(\boldsymbol{X}_i) \delta_k(Y_i,\, T_i,\, \boldsymbol{X}_i) + o_p(1).$$

for appropriate $\delta_k(Y_i,\, T_i,\, \boldsymbol{X}_i)$. Furthermore, using standard results for MLEs ([**?**]), it is not difficult to see that

$$\widehat{p}_k(\boldsymbol{x}) - p_k(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n S_k(T_i,\, \boldsymbol{X}_i;\, \boldsymbol{x}) + o_p(n^{-1})$$

where $S_k(T_i,\, \boldsymbol{X}_i;\, \boldsymbol{x})$ is a linear transformation of the score vector function for ith observation. Note that $E[S_k(T_i,\, \boldsymbol{X}_i;\, \boldsymbol{x})] = 0$. From the above relationships, it follows that

$$
\begin{aligned}
\widehat{\xi}_k - \xi_k &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_1}\left\{p_1(\boldsymbol{X}_i) + \frac{1}{n}\sum_{j=1}^{n} S_1(T_j,\, \boldsymbol{X}_j;\, \boldsymbol{X}_i)\right\}\delta_k(Y_i,\, T_i,\, \boldsymbol{X}_i) + o_p(n^{-1/2}) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{\delta}_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, T_j,\, \boldsymbol{X}_j) + o_p(n^{-1/2}) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\widetilde{\delta}_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, T_j,\, \boldsymbol{X}_j) + o_p(n^{-1/2}) \qquad (10)
\end{aligned}
$$

where

$$
\widetilde{\delta}_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, T_j,\, \boldsymbol{X}_j) = \{p_1(\boldsymbol{X}_i) + S_1(T_j,\, \boldsymbol{X}_j;\, \boldsymbol{X}_i)\}\,\delta_k(Y_i,\, T_i,\, \boldsymbol{X}_i).
$$

Again, the terms $\widetilde{\delta}_k(Y_i,\, T_i,\, \boldsymbol{X}_i,\, T_j,\, \boldsymbol{X}_j)$ have null expectation.

Define next the symmetric kernel

$$
h_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, Y_j,\, T_j,\, \boldsymbol{X}_j) = \frac{1}{2}\left(\widetilde{\delta}_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, T_j,\, \boldsymbol{X}_j) + \widetilde{\delta}_k(Y_j,\, T_j,\, \boldsymbol{X}_j;\, T_i,\, \boldsymbol{X}_i)\right)
$$

and consider the U-statistic of degree 2

$$
U_{n,k} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} h_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, Y_j,\, T_j,\, \boldsymbol{X}_j), \quad k = 0,\, 1. \qquad (11)
$$

Since $E[h_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, Y_j,\, T_j,\, \boldsymbol{X}_j)] = 0$, it is not difficult to see that $E[U_{n,k}] = \xi_k$. Define further

$$
\boldsymbol{U}_n = \begin{bmatrix} U_{n,0} \\ U_{n,1} \end{bmatrix}.
$$

From (10) it is seen that

$$
\sqrt{n}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}) = \sqrt{n}(\boldsymbol{U}_n - \boldsymbol{\xi}) + o_P(1). \qquad (12)
$$

Hence, from standard theory of U-statistics (cfr. [**?**], p. 192), it is easy to argue that $\sqrt{n}(\boldsymbol{U}_n - \boldsymbol{\xi})$ possesses the same limiting distribution as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i)$$

where

$$\boldsymbol{q}^*(Y_i,\, T_i,\, \boldsymbol{X}_i) = 2E[h_k(Y_i,\, T_i,\, \boldsymbol{X}_i;\, Y_j,\, T_j,\, \boldsymbol{X}_j)|Y_i,\, T_i,\, \boldsymbol{X}_i].$$

If $E[\boldsymbol{q}_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^2] > 0,\; k = 0,\, 1$, then statement 2. follows. $\square$

The asymptotic normality of ATT can be obtained similarly to that of WATE. Let $\boldsymbol{a}$ be the vector with components $1, -1$, and define

$$\widehat{\Delta}_{ATT} = \widehat{\xi}_1 - \widehat{\xi}_0.$$

Proposition B.2. *Suppose that assumptions A1-A3 are fulfilled, and that* $E[\boldsymbol{q}_k^*(Y_i,\, T_i,\, \boldsymbol{X}_i)^2] > 0,\; k = 0,\, 1$. *Then:*

$$\sqrt{n}(\widehat{\Delta}_{ATT} - \Delta_{ATT}) \xrightarrow{d} N(0,\, \sigma_a^2) \text{ as } n \to \infty$$

where $\sigma_a^2 = \boldsymbol{a}^T \Psi \boldsymbol{a}$.