Setup

Consider the following setup. Let $(X, Y_1, Y_2, \delta) \overset{ind}{\sim} F$ for some distribution F that is unknown. We define $Z = (X, Y_1, Y_2$ and we observe the following:

Table 1: This table shows some of our notation and some of the corresponding notation from [1].

Segments	X	Y_1	Y_2	Prob. Element in Segment	δ	C	$G_C(Z)$
A_{00}	✓			π_{00}	δ_{00}	1	$\{X\}$
A_{10}	✓	✓		π_{10}	δ_{10}	2	$\{X, Y_1\}$
A_{01}	✓		✓	π_{01}	δ_{01}	3	$\{X, Y_2\}$
A_{11}	✓	✓	✓	π_{11}	δ_{11}	∞	$\{X, Y_1, Y_2\}$

Define $\varpi(r, Z) = \Pr(C = r \mid Z)$. For now, we assume that $\varpi(r, Z)$ is known. Notice that $\varpi(\infty, Z) = \pi_1 1$, $\varpi(3, Z) = \pi_{01}$, $\varpi(2, Z) = \pi_{10}$, and $\varpi(1, Z) = \pi_{00}$. Since $\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$, we only need to define three inclusion probabilities.

Suppose that we want to estimate $\theta = E[g(X, Y_1, Y_2)]$ for a known function g. The goal is the show that over all functions $b_1(X, Y_1)$ and $b_2(X, Y_2)$, we have the optimal estimator in the form:

$$\hat{\theta} = \tag{1}$$
$$n^{-1}\sum_{i=1}^n E[g_i \mid X_i] + n^{-1}\sum_{i=1}^n \frac{\delta_{10}}{\pi_{10}}(b_1(X_i, Y_{1i}) - E[g_i \mid X_i]) + n^{-1}\sum_{i=1}^n \frac{\delta_{01}}{\pi_{01}}(b_2(X_i, Y_{2i}) - E[g_i \mid X_i])$$
$$+ n^{-1}\sum_{i=1}^n \frac{\delta_{11}}{\pi_{11}}(g_i - b_1(X_i, Y_{1i}) - b_2(X_i, Y_{2i}) + E[g_i \mid X_i])$$

Semiparametric Inference

We know from Theorem 7.2 of [1] that if $\Pr(C = \infty \mid Z) = \pi_{11} > 0$ then the semiparametric influence function has the form (see page 20 of notes):

$$\frac{I(C = \infty)g(Z)}{\varpi(\infty, Z)} + \frac{I(C = \infty)}{\varpi(\infty, Z)}\left(\sum_{r \neq \infty} \varpi(r, G_r(Z)) L_{2r}(G_r(Z))\right) - \sum_{r \neq \infty} I(C = r) L_{2r}(G_r(Z)) \tag{2}$$

where $L_{2r}(G_r(Z))$ is an arbitrary function of $G_r(Z)$. Notice, that this form does not identify *the* optimal estimator but a class of semiparametric functions. A reasonable choice of an estimator for $L_{2r}(G_r(Z))$ is

$$L_{2r}(G_r(Z)) = c_r E[g(Z) \mid G_r(Z)]$$

where each c_r is a constant. This actually suggests a variety of estimators that each have different values of coefficients of δ. Initially, we tried a direct projection onto the nuisance tangent space, but I got stuck. See Appendix A for this work.

Different Classes of Estimators

All of the considered estimators have a similar form:

$$\hat{\theta} = \frac{\delta_{11}}{\pi_{11}} g(Z) + \beta_0(\delta, c_0) E[g(Z) \mid X] + \beta_1(\delta, c_1) E[g(Z) \mid X, Y_1] + \beta_2(\delta, c_2) E[g(Z) \mid X, Y_2].$$

In this case, c is a vector of constants $c = (c_0, c_1, c_2)$ that are chosen to minimize the variance of $\hat{\theta}$. See Appendix B for details on an example about choosing the values of $\hat{c}$.

Estimator	$\beta_0(\delta, c_0)$	$\beta_1(\delta, c_1)$	Implemented
$\hat{\theta}_{prop}$	$\left(1 - \frac{(\delta_{10}+\delta_{11})}{(\pi_{10}+\pi_{11})} - \frac{(\delta_{01}+\delta_{11})}{(\pi_{01}+\pi_{11})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$\left(\frac{\delta_{10}+\delta_{11}}{\pi_{10}+\pi_{11}} - \frac{\delta_{11}}{\pi_{11}}\right)$	✓
$\hat{\theta}_{prop}^{ind}$	$\left(1 - \frac{(\delta_{10})}{(\pi_{10})} - \frac{(\delta_{01})}{(\pi_{01})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$\left(\frac{\delta_{10}}{\pi_{10}} - \frac{\delta_{11}}{\pi_{11}}\right)$	✓
$\hat{\theta}_c$	$c_0\left(1 - \frac{(\delta_{10}+\delta_{11})}{(\pi_{10}+\pi_{11})} - \frac{(\delta_{01}+\delta_{11})}{(\pi_{01}+\pi_{11})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$c_1\left(\frac{\delta_{10}+\delta_{11}}{\pi_{10}+\pi_{11}} - \frac{\delta_{11}}{\pi_{11}}\right)$	✓
$\hat{\theta}_{c,ind}$	$c_0\left(1 - \frac{(\delta_{10})}{(\pi_{10})} - \frac{(\delta_{01})}{(\pi_{01})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$c_1\left(\frac{\delta_{10}}{\pi_{10}} - \frac{\delta_{11}}{\pi_{11}}\right)$	✓
$\hat{\theta}_\delta$	$c_0\left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{00}}{\pi_{00}}\right)$	$c_1\left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{10}}{\pi_{10}}\right)$	✓

I wanted to give a brief comment on the origin of each of these estimators. The first estimator $\hat{\theta}_{prop}$ is the estimator proposed by Dr. Kim. The original goal was to show that this is optimal. The second estimator, $\hat{\theta}_{prop}^{ind}$ is the first estimator with different functions of δ to try to make each component more independent from each other. The problem with this is that the components are still dependent and it has worse properties in simulation studies (next section). The estimator $\hat{\theta}_c$ is the originial estimator $\hat{\theta}_{prop}$ with different coefficients that hopefully make it slightly more efficient. The estimator $\hat{\theta}_{c,ind}$ does the same thing with $\hat{\theta}_{prop}^{ind}$. Finally, $\hat{\theta}_\delta$ uses the functional form of Theorem 7.2 from [1] and optimized coefficients.

Simulation Studies

Simulation 1

To see which of these estimators is the best, we run a simulation study. In this study, we have the following variables:

$$\begin{bmatrix} x \\ e_1 \\ e_2 \end{bmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}\right)$$
$$y_1 = x + e_1$$
$$y_2 = \theta + x + e_2$$

Furthermore, $\pi_{11} = 0.4$ and $\pi_{00} = \pi_{10} = \pi_{01} = 0.2$. The goal of this simulation study is to find $\theta = E[Y_2]$. In other words, $g(Z) = Y_2$. There are several algorithms for comparison which are defined as the following:

$$Oracle = n^{-1} \sum_{i=1}^{n} g(Z_i)$$
$$CC = \frac{\sum_{i=1}^{n} \delta_{11} g(Z_i)}{\sum_{i=1}^{n} \delta_{11}}$$
$$IPW = \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g(Z_i)$$

Table 2: The true value of θ is 5. $\rho = 0.5$. The test statistic and p-value are from a one-sample t-test to see if the estimator is biased.

Algorithm	Bias	SD	Tstat	Pval
Oracle	0.001	0.044	1.546	0.061
CC	0.002	0.058	1.549	0.061
IPW	0.006	0.210	1.504	0.066
$\hat{\theta}_{prop}$	0.002	0.051	1.911	0.028
$\hat{\theta}_c$	0.002	0.051	1.920	0.027
$\hat{\theta}_{prop}^{ind}$	0.002	0.091	0.983	0.163
$\hat{\theta}_{c,ind}$	0.001	0.074	0.784	0.216
$\hat{\theta}_\delta$	0.002	0.051	1.964	0.025

The main takeaway from this simulation study is that $\hat{\theta}_\delta$, $\hat{\theta}_c$ and $\hat{\theta}_{prop}$ have the smallest standard deviation among the non-oracle estimators. In fact $\hat{\theta}_c$ and $\hat{\theta}_{prop}$ are virtually identical because the optimal values of c in $\hat{\theta}_c$ are $(c_0, c_1, c_2)' \approx (1, 1, 1)'$.

Simulation 2

We now test a second simulation study. The setup of this simulation study is the same as the previous simulation. The only difference is that we are interested in estimating $\theta = E[Y_1^2 Y_2]$. In other words, now we have $g(Z) = Y_1^2 Y_2$.

Table 3: The true value of $E[g(Z)]$ is 10. $\rho = 0.5$. The test statistic and p-value are from a one-sample t-test to see if the estimator is biased.

Algorithm	Bias	SD	Tstat	Pval
Oracle	0.007	0.529	0.741	0.229
CC	0.023	0.824	1.518	0.065
IPW	0.031	0.915	1.846	0.032
$\hat{\theta}_{prop}$	0.011	0.635	0.912	0.181
$\hat{\theta}_c$	0.012	0.632	1.010	0.156
$\hat{\theta}_{c,ind}$	0.013	0.662	1.060	0.145
$\hat{\theta}_\delta$	0.011	0.636	0.983	0.163

Like the first simulation, $\hat{\theta}_{prop}$, $\hat{\theta}_c$ and $\hat{\theta}_\delta$ performed the best.

Next Steps

1. It turns out that [1] contains a general technique for solving non-monotone missing data. They show that the optimal estimator is found via the following double projection:

$$L_2 = \Pi\left(\Pi\left(\frac{\delta_{11}}{\pi_{11}} g(Z) \mid C, G_C(Z)\right) \mid Z\right)$$

I can show that $\hat{\theta}_\delta$ is NOT such a double projection. However, I think it is worth exploring what such an estimator (especially if we consider a linear estimator) looks like.

2. Similar to the previous point, [1] discuss what the optimal estimator of a non-monotone problem looks like and they determine that it solves the estimating equation

$$\sum_{i=1}^{n} \mathcal{L}(\mathcal{M}^{-1}(g(Z)))$$

where $\mathcal{M}(g(Z)) = E[\mathcal{L}(g(Z)) \mid Z]$ and $\mathcal{L}(g(Z)) = E[g(Z) \mid C, G_C(Z)]$. The difficult part is understanding $\mathcal{M}^{-1}$, but there is conveniently the fact that

$$\mathcal{M}^{-1}(g(Z)) = \lim_{n\to\infty} \phi_{n+1}(Z) \text{ where } \phi_{n+1}(Z) = (I - \mathcal{M})\phi_n(Z) + g(Z) \text{ and } \phi_0(Z) = g(Z).$$

I have been investigating what this estimating equation yields if we approximate $\mathcal{M}^{-1}$ with ϕ_1. However, I have not finished this yet.

3. A different method to find the projection onto Λ_2 could be to try to minimize the (KL) divergence between L_2 and $\frac{\delta_{11}}{\pi_{11}} g(Z)$.

References

[1] Anastasios A Tsiatis. “Semiparametric theory and missing data”. In: (2006).

A Linear Expectations

To simplify this problem, we consider the following estimator[1]:

$$
\begin{aligned}
&\hat{\theta}_c \qquad (3)\\
&= \frac{\delta_{11}}{\pi_{11}} g(Z) + \left(1 - \left(\frac{\delta_{10} + \delta_{11}}{\pi_{10} + \pi_{11}}\right) - \left(\frac{\delta_{01} + \delta_{11}}{\pi_{01} + \pi_{11}}\right) + \frac{\delta_{11}}{\pi_{11}}\right) c_0 E[g \mid X] \\
&+ \left(\left(\frac{\delta_{10} + \delta_{11}}{\pi_{10} + \pi_{11}}\right) - \frac{\delta_{11}}{\pi_{11}}\right) c_1 E[g \mid X, Y_1] + \left(\left(\frac{\delta_{01} + \delta_{11}}{\pi_{01} + \pi_{11}}\right) - \frac{\delta_{11}}{\pi_{11}}\right) c_2 E[g \mid X, Y_2]
\end{aligned}
$$

A.1 Projection onto Nuisance Tangent Space

The goal is now to find the coefficients c_0, c_1, and c_2 such that $\langle \hat{\theta}_c, L_2 \rangle \equiv E[\hat{\theta}_c L_2] = 0$ for all $L_2 \in \Lambda_2$ (see [1] for definition of Λ_2). If we can find such coefficients that the estimator $\hat{\theta}_c$ will be orthogonal to Λ_2 and hence by Theorem 10.1 of [1] semiparametrically optimal. The good news is that we know (from Theorem 7.2) that any element $L_2 \in \Lambda_2$ has a form:

$$
L_2 = \left(\frac{\delta_{11}}{\pi_{11}} \pi_{00} - \delta_{00}\right) L_{20}(X) + \left(\frac{\delta_{11}}{\pi_{11}} \pi_{10} - \delta_{10}\right) L_{21}(X, Y_1) + \left(\frac{\delta_{11}}{\pi_{11}} \pi_{01} - \delta_{01}\right) L_{22}(X, Y_2). \tag{4}
$$

Then expanding and solving $E[\hat{\theta}_c L_2] = 0$ yields:

$$
\begin{aligned}
0 &= E[\hat{\theta}_c L_2] \\
& E\Bigg[\left(\frac{\pi_{00}}{\pi_{11}} + \left(\frac{\pi_{10}}{\pi_{10} + \pi_{11}}\right) \frac{\pi_{00} c_1}{\pi_{11}} + \left(\frac{\pi_{01}}{\pi_{01} + \pi_{11}}\right) \frac{\pi_{00} c_2}{\pi_{11}} + \frac{\pi_{00}(\pi_{10}\pi_{01} - \pi_{11}^2) c_0}{(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})\pi_{11}}\right) E[g \mid X] L_{20}(X) \\
&+ \frac{\pi_{10}}{\pi_{11}} \left(E[g(Z) L_{21}(X, Y_1) \mid X] - c_1 E[E[g(Z) \mid X, Y_1] L_{21}(X, Y_1) \mid X] + \frac{\pi_{01} c_2}{\pi_{10} + \pi_{11}} E[E[g \mid X, Y_2] L_{21}(X, Y_1) \mid X] + \frac{\pi_{10}\pi_{01}}{\pi_{11}(\pi_{01} + \pi_{11})} E[g \mid X] E[L_{21}(X, Y_1) \mid X] c_0\right) \\
&+ \frac{\pi_{01}}{\pi_{11}} \left(E[g(Z) L_{22}(X, Y_2) \mid X] + \frac{\pi_{10} c_1}{\pi_{10} + \pi_{11}} E[E[g(Z) \mid X, Y_1] L_{22}(X, Y_2) \mid X] - c_2 E[E[g \mid X, Y_2] L_{22}(X, Y_2) \mid X] + \frac{\pi_{10}}{(\pi_{01} + \pi_{11})} E[g \mid X] E[L_{22}(X, Y_2) \mid X] c_0\right)\Bigg]
\end{aligned}
$$

To solve for c_0, c_1, and c_2 we need the following to hold for any $L_{21}(X, Y_1)$ and $L_{22}(X, Y_2)$:

$$
\begin{aligned}
&1 + c_0 \frac{\pi_{01}\pi_{10} - \pi_{11}^2}{(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})} + c_1 \frac{\pi_{10}}{\pi_{01} + \pi_{11}} + c_2 \frac{\pi_{01}}{\pi_{01} + \pi_{11}} = 0 \\
&E\left[\left(g(Z) + c_0 \frac{\pi_{01}}{\pi_{01} + \pi_{11}} E[g(Z) \mid X] - c_1 E[g(Z) \mid X, Y_1] + c_2 \frac{\pi_{01}}{\pi_{10} + \pi_{11}} E[g(Z) \mid X, Y_2]\right) L_{21}(X, Y_1) \mid X\right] = 0 \\
&E\left[\left(g(Z) + c_0 \frac{\pi_{10}}{\pi_{10} + \pi_{11}} E[g(Z) \mid X] + c_1 \frac{\pi_{10}}{\pi_{10} + \pi_{11}} E[g(Z) \mid X, Y_1] - c_2 E[g(Z) \mid X, Y_2]\right) L_{22}(X, Y_2) \mid X\right] = 0
\end{aligned}
$$

Unfortunately, I am now stuck because it is unclear to me how to use any actual data to solve this problem.

[1] This estimator has slightly different coefficients compared to the initial estimator.

B Solving for $\hat{c}$

We can find the values of c_0, c_1, and c_2 in $\hat{\theta}_c$ that minimize the variance the estimator. We can find these values by differentiating by c_i and solving for c_i:

$$\begin{bmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \hat{c}_2 \end{bmatrix} = - \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}^{-1} \times \begin{bmatrix} E[E[g \mid X]^2] \left(1 + \frac{\pi_{10}\pi_{01} - \pi_{11}^2}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}\right) \\ E[E[g \mid X, Y_1]^2] \left(\frac{-\pi_{10}}{\pi_{11}(\pi_{10}+\pi_{11})}\right) \\ E[E[g \mid X, Y_2]^2] \left(\frac{-\pi_{01}}{\pi_{11}(\pi_{01}+\pi_{11})}\right) \end{bmatrix}$$

where

$$M_{11} = E[E[g \mid X]^2] \left(\frac{\pi_{11}^2 + \pi_{10}\pi_{01}}{\pi_{11}(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})} - 1\right)$$
$$M_{12} = E[E[g \mid X]^2] \left(\frac{-\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})}\right)$$
$$M_{13} = E[E[g \mid X]^2] \left(\frac{-\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})}\right)$$
$$M_{22} = E[V(E[g \mid X, Y_1] \mid X)] \left(\frac{\pi_{10}}{\pi_{11}(\pi_{10} + \pi_{11})}\right)$$
$$M_{23} = E[E[g \mid X, Y_1]E[g \mid X, Y_2]] \left(\frac{\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})}\right)$$
$$M_{33} = E[V(E[g \mid X, Y_2] \mid X)] \left(\frac{\pi_{01}}{\pi_{11}(\pi_{01} + \pi_{11})}\right)$$

This estimator is similar to several other estimators:

$$\begin{aligned} \hat{\theta}_c^{ind} = & \frac{\delta_{11}}{\pi_{11}} g(Z) + \left(1 - \left(\frac{\delta_{10}}{\pi_{10}}\right) - \left(\frac{\delta_{01}}{\pi_{01}}\right) + \frac{\delta_{11}}{\pi_{11}}\right) c_0 E[g \mid X] \\ & + \left(\frac{\delta_{10}}{\pi_{10}} - \frac{\delta_{11}}{\pi_{11}}\right) c_1 E[g \mid X, Y_1] + \left(\frac{\delta_{01}}{\pi_{01}} - \frac{\delta_{11}}{\pi_{11}}\right) c_2 E[g \mid X, Y_2] \end{aligned} \tag{5}$$

$$\begin{aligned} \hat{\theta}_c^{\delta} = & \frac{\delta_{11}}{\pi_{11}} g(Z) + \left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{00}}{\pi_{00}}\right) c_0 E[g \mid X] \\ & + \left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{10}}{\pi_{10}}\right) c_1 E[g \mid X, Y_1] + \left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{01}}{\pi_{01}}\right) c_2 E[g \mid X, Y_2] \end{aligned} \tag{6}$$

The first expression (Equation 5) is the proposed estimator with independent differences in each segment, while the second expression (Equation 6) is the optimal estimator with values of δ such that $\hat{\theta}_c^{\delta} \in \Lambda_2$, which means that it has the form of the semiparametric in Equation 2.