The Efficiency Loss of Semiparametric Inference

Inspired by Dr. Fuller's note, I set out to see if we could assess the amount of efficiency loss due to using semiparametric inference. First, let's define some notation.

Segment	Quantity of Interest	Estimator	Coefficient
A_{00}	$E[g \mid X]$	$\hat{\gamma}_{11}$	c_{11}
A_{10}	$E[g \mid X]$	$\hat{\gamma}_{21}$	c_{21}
A_{10}	$E[g \mid X, Y_1]$	$\hat{\gamma}_{22}$	c_{22}
A_{01}	$E[g \mid X]$	$\hat{\gamma}_{31}$	c_{31}
A_{01}	$E[g \mid X, Y_2]$	$\hat{\gamma}_{33}$	c_{33}
A_{11}	$E[g \mid X]$	$\hat{\gamma}_{41}$	c_{41}
A_{11}	$E[g \mid X, Y_1]$	$\hat{\gamma}_{42}$	c_{42}
A_{11}	$E[g \mid X, Y_2]$	$\hat{\gamma}_{43}$	c_{43}
A_{11}	$E[g \mid X, Y_1, Y_2]$	$\hat{\gamma}_{44}$	c_{44}

Table 1: This table shows the relationship between different quantities of interest in each segment, their estimators, and the corresponding coefficients of their estimators.

This is a similar framework to understanding the class of estimators that we discussed in `optest_update.pdf`, where we write the estimator as a weighted average of different expectation with functional coefficients.

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

However, now we write the estimator as

$$\hat{\theta} = \sum_{k,t} c_{kt} \hat{\gamma}_{kt}$$

where $\hat{\gamma}_{kt} = \frac{\delta_{ij}}{\pi_{ij}} E[g \mid G_{ij}(X, Y_1, Y_2)]$ and ij corresponds to the segment A_{ij} associated with $\hat{\gamma}_{kt}$ in Table 1.

Different Classes of Models

As noted by Dr. Fuller, we can understand parametric estimation as solving the following constrained optimization problems:

$$\text{minVar}\left(\sum_{k,t} c_{kt}\hat{\gamma}_{kt}\right) \text{ such that } \sum_{k,t} c_{kt} = 1$$

or

$$\text{minVar}\left(\sum_{k,t} c_{kt}\hat{\gamma}_{kt}\right) \text{ such that } \sum_{(k,t):(k,t)\neq(4,4)} c_{kt} = 0 \text{ and } c_{44} = 1.$$

To be robust to the outcome model, we can construct a similar optimization problem with different constraints:

$$\min \text{Var}\left(\sum_{k,t} c_{kt}\hat{\gamma}_{kt}\right) \text{ such that } c_{11} + c_{21} + c_{31} + c_{41} = 0, c_{22} + c_{42} = 0, c_{33} + c_{43} = 0, \text{ and } c_{44} = 1.$$

Likewise to be robust to the response model, we can have the problem:

$$\min \text{Var}\left(\sum_{k,t} c_{kt}\hat{\gamma}_{kt}\right) \text{ such that } c_{11} = \pi_{00}, c_{21} + c_{22} = \pi_{10}, c_{31} + c_{33} = \pi_{01}, \text{ and } c_{41} + c_{42} + c_{43} + c_{44} = \pi_{11}.$$

To be double robust (robust to the outcome and response model) we can combine the last two constraints. This is summarized in Table 2.

Table 2: This table identifies the different constraints for each model type.

Type	Constraints
Parametric 1	$\sum_{k,t} c_{kt} = 1$
Parametric 2	$\sum_{k,t:(k,t)\neq(4,4)} c_{kt} = 0, c_{44} = 1$
Outcome Robust	$c_{11} + c_{21} + c_{31} + c_{41} = 0, c_{22} + c_{42} = 0, c_{33} + c_{43} = 0,$ and $c_{44} = 1.$
Response Robust	$c_{11} = \pi_{00}, c_{21}+c_{22} = \pi_{10}, c_{31}+c_{33} = \pi_{01},$ and $c_{41}+c_{42}+c_{43}+c_{44} = \pi_{11}$
Double Robust	$c_{11} + c_{21} + c_{31} + c_{41} = 0, c_{22} + c_{42} = 0, c_{33} + c_{43} = 0, c_{44} = 1,$ $c_{11} = \pi_{00}, c_{21}+c_{22} = \pi_{10}, c_{31}+c_{33} = \pi_{01},$ and $c_{41}+c_{42}+c_{43}+c_{44} = \pi_{11}$

Simulation Study

Simulation 1

We use the following simulation setup

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

Furthermore, $\pi_{11} = 0.2$ and $\pi_{00} = 0.3, \pi_{10} = 0.4$, and $\pi_{01} = 0.1$. 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:

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

Furthermore, we include four existing estimators that we have already proposed in the past: WLS, Prop, PropInd, and SemiDelta.

The estimator WLS is a a weight least square estimator derived in the following manner. We now consider a normal model:

$$\begin{pmatrix} x_i \\ e_{1i} \\ e_{2i} \end{pmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_{11} & \sigma_{12} \\ 0 & \sigma_{12} & \sigma_{22} \end{bmatrix}, \right)$$

and define $y_{1i} = \theta_1 + x_i + e_{1i}$ and $y_{2i} = \theta_2 + x_i + e_{2i}$. Then $b_1 = b_2 = 1$. We define $\bar{z}_k^{(ij)}$ as the mean of y_k in segment A_{ij}. This means that we have means $\bar{z}_1^{(11)}$, $\bar{z}_2^{(11)}$, $\bar{z}_1^{(10)}$, and $\bar{z}_2^{(01)}$. Let $W = [\bar{z}_1^{(11)}, \bar{z}_2^{(11)}, \bar{z}_1^{(10)}, \bar{z}_2^{(01)}]'$, then for $n_{ij} = |A_{ij}|$, we have

$$Z - M\mu \sim N(\vec{0}, V)$$

where

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } V = \begin{bmatrix} \frac{\sigma_{11}}{n_{11}} & \frac{\sigma_{12}}{n_{11}} & 0 & 0 \\ \frac{\sigma_{12}}{n_{11}} & \frac{\sigma_{22}}{n_{11}} & 0 & 0 \\ 0 & 0 & \frac{\sigma_{11}}{n_{10}} & 0 \\ 0 & 0 & 0 & \frac{\sigma_{22}}{n_{01}} \end{bmatrix}.$$

Thus, the BLUE for $\mu = [\mu_1, \mu_2]'$ is

$$\hat{\mu} = (M'V^{-1}M)^{-1}M'V^{-1}W.$$

Hence, WLS is μ_2 as $g(X, Y_1, Y_2) = Y_2$.

The remaining three estimators are drived from the following expression where the values for β are provided in Table 3. These are good estimators to compare to because Prop is the original proposed estimator. PropInd has the same form as Prop except the values for β is different. PropInd shares the same values of β as all of the new models with constraints. SemiDelta is useful because it is the best estimator in general so far.

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

Table 3: This table displays the values of β for different estimator types.

Estimator	$\beta_0(\delta, c_0)$	$\beta_1(\delta, c_1)$
Prop	$\left(1 - \frac{(\delta_{10}+\delta_{11})}{(\pi_{10}+\pi_{11})} - \frac{(\delta_{01}+\delta_{11})}{(\pi_{01}+\pi_{11})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$\left(\frac{\delta_{10}+\delta_{11}}{\pi_{10}+\pi_{11}} - \frac{\delta_{11}}{\pi_{11}}\right)$
PropInd	$\left(1 - \frac{(\delta_{10})}{(\pi_{10})} - \frac{(\delta_{01})}{(\pi_{01})} + \frac{\delta_{11}}{\pi_{11}}\right)$	$\left(\frac{\delta_{10}}{\pi_{10}} - \frac{\delta_{11}}{\pi_{11}}\right)$
SemiDelta	$c_0\left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{00}}{\pi_{00}}\right)$	$c_1\left(\frac{\delta_{11}}{\pi_{11}} - \frac{\delta_{10}}{\pi_{10}}\right)$

In the simulation results in Table 4, the new results have the same label as the value from the "Type" column in Table 2.

Table 4: Results from Simulation 1. True values: $\theta = 5, \rho = 0.5$. The test conducted for the T-statistic and P-value is a two sample test to see if the estimator is unbiased.

Algorithm	Bias	SD	Tstat	P-value
Oracle	-0.002	0.045	-1.669	0.048
CC	0.001	0.083	0.263	0.396
IPW	0.010	0.357	0.872	0.192
WLS	0.000	0.054	-0.065	0.474
Prop	-0.002	0.063	-0.768	0.221
PropInd	-0.003	0.120	-0.900	0.184
SemiDelta	-0.001	0.064	-0.342	0.366
Parametric 1	-0.001	0.061	-0.509	0.305
Parametric 2	-0.001	0.061	-0.509	0.305
Outcome Robust	-0.001	0.061	-0.543	0.294
Response Robust	-0.001	0.061	-0.509	0.305
Double Robust	-0.001	0.061	-0.543	0.294

Simulation 2

Next, we used the same simulation setup except that we are interested in estimating $\theta = E[g] = E[Y_1^2 Y_2]$. We do not use the WLS estimator because it does not work.

Table 5: Results from Simulation 2. True values: $\theta = 10, \rho = 0.5$. The test conducted for the T-statistic and P-value is a two sample test to see if the estimator is unbiased.

Algorithm	Bias	SD	Tstat	P-value
Oracle	-0.037	0.528	-2.214	0.014
CC	0.007	1.196	0.179	0.429
IPW	0.023	1.405	0.520	0.302
Prop	-0.044	0.691	-2.027	0.021
SemiDelta	-0.032	0.674	-1.521	0.064
Parametric 1	-0.004	0.301	-0.460	0.323
Parametric 2	-0.004	0.344	-0.341	0.367
Outcome Robust	-0.040	0.654	-1.950	0.026
Response Robust	-0.004	0.301	-0.461	0.322
Double Robust	-0.040	0.654	-1.954	0.025

Discussion

- In the first simulation there does not seem to be much of a cost to using a semiparametric estimator with the Double Robust model performing as well as the Outcome Robust, Response Robust and Parametric models. However, in the second simulation the Parametric models and Response Robust estimators perform the best.
- Strangly, the Double Robust estimator in the second simulation has picked up the bad characteristics of the Outcome Model. This can make sense if the optimal Outcome Model is contained in the space of the Double Robust model but the optimal Response Model is not.