1 Comments on the Simulation Program

For each of the designs there are n units in the population. For notational brevity, we omit the variable subscript n everywhere in this note, although it remains implicit that all variables are generated with respect to population n.

In this simulation study, we focus on a single causal variable U_i. Each unit i is characterized by three variables, ξ_i, θ_i and Z_i. The non-stochastic pair (ξ_i, θ_i) summarizes unit i's potential outcome dependence on observed and unobserved attributes. In our setting, the fixed observed attributes are included in Z_i, a $(KZ + 1)$-dimensional vector with the first element equal to 1, and the remaining KZ elements drawn from a normal distribution with mean zero and unit variance, all independent. The population is generated first, with:

$$Z_{i1} = 1, \qquad Z_{ik} \sim \mathcal{N}(0, 1).$$

$$\xi_i \sim \mathcal{N}(Z_i^\top \gamma, \sigma_\xi^2)$$

$$\theta_i \sim \mathcal{N}(Z_i^\top \psi, \sigma_\theta^2).$$

In this particular simulation study, we assume all throughout $\sigma_\xi = 1$ such that $\xi_i \sim \mathcal{N}(Z_i^\top \gamma, 1)$. Having generated our population of interest, the potential outcome function for unit i with $U_i = u$ is given by:

$$Y_i(u) = \theta_i u + \xi_i.$$

The assignment mechanism is

$$U_i \sim \mathcal{N}(Z_i^\top \lambda, 1).$$

We work with the potential outcome function in terms of the causal variable X_i which is free of the correlation from Z_i. Since for our simulations, we have $\mathbb{E}[U_i] = 0$ (by (3.1) in the paper) this implies that $\lambda = 0$, such that it holds $X_i = U_i$, and we can directly re-write the potential outcome function as [1]:

$$Y_i(u) = \theta_i x + \xi_i.$$

Once we have this population, we can calculate the various estimands and the various variances. Because X_i and Z_i are uncorrelated, for the limit of the estimands we have:

$$\lim_{n\to\infty} \theta_n^{causal} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \theta_i = \mathbb{E}[Z]^\top \psi = \psi_0 = 0$$

[1]The way this is implemented in the code is by the direct assumption that U_i and Z_i are uncorrelated, i.e., by directly imposing $\lambda = 0$ rather than by assuming $\mathbb{E}[U_i] = 0$.

in all the designs. Furthermore,

$$\lim_{n\to\infty} \gamma_n = \mathbb{E}[Z_i Z_i^\top]^{-1}\mathbb{E}[Z_i Y_i] = \mathbb{E}[Z_i Y_i].$$

For the first element of the limit of γ_n, corresponding to the intercept, we have:

$$\lim_{n\to\infty} \gamma_{1n} = \mathbb{E}[Y_i] = \mathbb{E}[\xi_i] + \mathbb{E}[\theta_i X_i] = 0 + \mathbb{E}[Z_i^\top \psi_i X_i] + \mathbb{E}[(\theta_i - Z_i^\top \psi)X_i] = 0$$

For the other elements:

$$\lim_{n\to\infty} \gamma_{kn} = \mathbb{E}[Z_{ik} Y_i] = \mathbb{E}[Z_{ik}\xi_i] + \mathbb{E}[Z_{ik}\theta_i X_i]$$

$$= \mathbb{E}[Z_{ik} Z_i^\top \gamma] + \mathbb{E}[Z_{ik}(\xi_i - Z_i^\top \gamma)] + \mathbb{E}[Z_{ik} Z_i^\top \psi_i X_i] + \mathbb{E}[Z_{ik}(\theta_i - Z_i^\top \psi)X_i] = 0$$

Hence, in the simulations we set $\gamma = 0$ which makes sense for large populations. Furthermore, to calculate θ_n^{causal}, notice that the formula in (3.5) simplifies to Ω_n^{XY} since both X, Z are standard normal and uncorrelated such that the first matrix is just the identity. Hence:

$$\Omega_n^{XY} = \frac{1}{n}\mathbb{E}[Y_i X_i] = \frac{1}{n}\mathbb{E}[x^2\theta_i + x\xi_i] = \frac{1}{n}\theta_i \mathbb{E}[x^2] = \frac{1}{n}\theta_i$$

where we notice $\mathbb{E}[x\xi_i] = 0$ by construction, and $\mathbb{E}[x^2] = Var[x] = 1$ since X_i is standard normal. Similarly, we can derive that $\theta_n^{causal,sample} = \frac{1}{\sum_i R_i}\theta_i$ where R_i is the sampling indicator.