1 Inputs

- Data $\{Y_i, X_i, D_i, Z_i\}_{i=1}^n$
  - $Y_i \in \mathbb{R}$: outcome
  - $X_i \in \mathbb{R}^p$: individual covariates
    Partition X_i into a vector of continuous variables $X_{ci} \in \mathbb{R}^{p_c}$ and a vector of discrete/categorical variables $X_{di} \in \mathbb{R}^{p_d}$, so that $X_i = (X_{ci}, X_{di})$ and $p = p_c + p_d$.
    (We may want to either feed X_{ci} and X_{di} as two separate inputs, or feed X_i and a vector containing the indices of continuous variables in X_i)
  - $D_i \in \{0, 1\}$: binary treatment assignment
  - $Z_i \in \{0, 1\}$: binary treatment recommendation
- $ML : \mathbb{R}^p \to [0, 1]$: ML is a function that maps a covariate vector into the probability that $Z_i = 1$
- $\delta > 0$: bandwidth
- S: the number of simulation draws

2 Estimation Steps

1. Let C denote the set of indices of continuous variables in X_i. For each continuous variable X_{ik}, $k \in C$, calculate the mean μ_k and standard deviation σ_k, and standardize the variable by replacing X_{ik} with $(X_{ik} - \mu_k)/\sigma_k$.
2. For each $i = 1, ..., n$, compute QPS $p^s(X_i; \delta)$ as follows.
   (a) Draw S vectors $X_c^{(1)}, \ldots, X_c^{(S)} \in \mathbb{R}^{p_c}$ of continuous variables independently from the uniform distribution on $N(X_{ci}, \delta)$, where $N(X_{ci}, \delta) \equiv \{x_c \in \mathbb{R}^{p_c} : \|X_{ci} - x_c\| \leq \delta\}$ ($\|\cdot\|$ denotes the Euclidean distance), namely, $N(X_{ci}, \delta)$ is the p_c-dimensional ball centered at X_{ci} with radius δ.
   (See Appendix B.4 of our draft for how to sample from uniform distribution on a multi-dimensional ball. I think that Methods 2 & 3 are easy and efficient.)
   (b) Bring $X_c^{(1)}, \ldots, X_c^{(S)}$ back before standardization by replacing $X_k^{(s)}$ with $\sigma_k X_k^{(s)} + \mu_k$ for each $k \in C$.
   (c) Calculate
   $$p^s(X_i; \delta) = \frac{1}{S} \sum_{s=1}^{S} ML(X_c^{(s)}, X_{di}).$$
   Here we input the simulation draws $X_c^{(1)}, \ldots, X_c^{(S)}$ of continuous variables while holding the discrete variables X_{di} fixed at i's value.

   We need to compute $p^s(X_i; \delta)$ independently across i, so the total number of simulation draws will be $n \times S$.

3. Using the observations with $p^s(X_i;\delta) \in (0,1)$ only, run the following 2SLS regression:

$$\begin{aligned} \text{1st Stage:} \quad & D_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 p^s(X_i;\delta) + \nu_i \\ \text{2nd Stage:} \quad & Y_i = \beta_0 + \beta_1 D_i + \beta_2 p^s(X_i;\delta) + \epsilon_i. \end{aligned}$$

Let $W_i = \begin{bmatrix} 1 \\ Z_i \\ p^s(X_i;\delta) \end{bmatrix}$ and $V_i = \begin{bmatrix} 1 \\ D_i \\ p^s(X_i;\delta) \end{bmatrix}$. The 2SLS estimator $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$ is given by

$$\hat{\beta} = (\sum_{i=1}^{n} W_i V_i')^{-1} \sum_{i=1}^{n} W_i Y_i.$$

For calculating standard errors, use the heteroskedasticity robust variance estimator given by

$$\widehat{\text{Var}}(\hat{\beta}) = (\sum_{i=1}^{n} W_i V_i')^{-1} (\sum_{i=1}^{n} \hat{\epsilon}_i^2 W_i W_i') (\sum_{i=1}^{n} V_i W_i')^{-1},$$

where $\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \hat{\beta}_2 p^s(X_i;\delta)$.

3 Outputs

- The estimate $\hat{\beta}$ and its variance-covariance matrix $\widehat{\text{Var}}(\hat{\beta})$
- The number of observations with $p^s(X_i;\delta) \in (0,1)$