Simulation Scheme

The simulation study contains two parts

- Y is continuous
- Y is binary

Continuous

For the first type (i.e., Y is continuous), we consider the following two cases

- Linear regression
- Nonlinear regression (tentatively)

The linear regression setting use Gaussian error terms, and we consider both the correct model and the misspecified model. The label shift assumption is that $p(\mathbf{x}|y)$ remains the same, however, deriving the explicit form of $p(\mathbf{x}|y)$ can be difficult. For example, for the following model

$$Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_1X_3 + \epsilon,$$

where $\epsilon \sim \text{Norm}(0, \sigma^2)$. Even if $\mathbf{X}$ has a multivariate normal distribution, deriving the closed form expression for $p(\mathbf{x}|y)$ is non-trivial. Thus, we used the rejection sampling algorithm to generate samples. We need to find a constant $C_{\max}$ such that $\rho(y; \boldsymbol{\beta})/C_{\max}$ is bounded by 1. Here, we outline the procedure to generate samples from the target distribution as follows,

1. Generate a sample of $\mathbf{X} = \mathbf{x}$ from the marginal distribution of $p_s(\mathbf{x})$.
2. Given the sample $\mathbf{X} = \mathbf{x}$, generate $Y = y$ using the specified model $p_s(y|\mathbf{x})$.
3. Generate a number $U = u$ from the uniform distribution $\text{U}(0, 1)$.
4. If $U = u < \rho(y; \boldsymbol{\beta})/C_{\max}$, accept this pair $(\mathbf{x}, y)$; otherwise reject this pair.
5. Repeat the previous steps until accumulate the desired sample size.

Setting 1

The marginal distribution of $\mathbf{X} = (X_1, X_2, X_3)^{\text{T}}$ is

$$\mathbf{X} \sim \text{MVN}\left(\mu, \Sigma\right),$$

where

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The model $p_s(y|\mathbf{x})$ is

$$Y = 1 + 2X_1 + 2X_2 + 2X_3 + X_1X_2 + X_2X_3 + \epsilon,$$

where $\epsilon \sim \text{Norm}(0, 1)$.

References