Adversarial Estimation of Heterogeneous Treatment Effects

1 Heterogenous treatment effect

Consider a heterogeneous treatment effect model,

$$Y_i = \tau_i d_i + f\left(\boldsymbol{X}_i\right) + U_i,\ i \in \mathcal{N}, \tag{1.1}$$

where the treatment d_i is assigned in a random experiment, and $\boldsymbol{X}_i \in \mathbb{R}^p$ is the observed characteristics. Both treated and control units are drawn from the same super population. Denote $\mathcal{T} = \{i \in \mathcal{N} : d_i = 1\}$ and $\mathcal{C} = \mathcal{N} \setminus \mathcal{T}$ as the sets of treated and control units, respectively, and correspondingly $N_1 = |\mathcal{T}|$ and $N_0 = |\mathcal{C}|$.

Suppose $\{\tau_i\}_{i\in\mathcal{T}}$ is known, define

$$\tilde{Y}_i = \begin{cases} Y_i & \text{if } d_i = 0, \\ Y_i - \tau_i & \text{if } d_i = 1, \end{cases} \quad i \in \mathcal{N}, \tag{1.2}$$

then

$$\tilde{Y}_i = f\left(\boldsymbol{X}_i\right) + U_i, \forall i \in \mathcal{N},$$

i.e. $S_{\mathcal{T}} = \left\{\tilde{Y}_i, \boldsymbol{X}_i\right\}_{i\in\mathcal{T}}$ and $S_{\mathcal{C}} = \left\{\tilde{Y}_i, \boldsymbol{X}_i\right\}_{i\in\mathcal{C}}$ follow the same data generating process. In this case, one cannot $S_{\mathcal{T}}$ and $S_{\mathcal{C}}$.

In practice, the heterogeneous treatment effects τ_i are unknown parameter of interests. In Wager and Athey (2018), τ_i is modeled,

$$\tau\left(\boldsymbol{x}\right) = \mathbb{E}\left(\tau_i | \boldsymbol{X}_i = \boldsymbol{x}\right).$$

Following the intuition from the case where $\{\tau_i\}_{i\in\mathcal{T}}$ is known, we propose to adopt the generative adversarial network (GAN) framework (Goodfellow et al., 2014; Kaji et al., 2022) to estimate $\{\tau_i\}_{i\in\mathcal{T}}$. Consider a minimax game between two components, a generator G and a discriminator D , which can be modeled as deep neural networks. The estimation problem is defined as

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} \frac{1}{N_1} \sum_{i\in\mathcal{T}} \log D\left(\tilde{Y}_i\left(G\left(\boldsymbol{X}_i\right)\right), \boldsymbol{X}_i\right) + \frac{1}{N_0} \sum_{i\in\mathcal{C}} \log\left(1 - D\left(Y_i, \boldsymbol{X}_i\right)\right). \tag{1.3}$$

References

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 27. Curran Associates, Inc.

Kaji, T., E. Manresa, and G. Pouliot (2022). An adversarial approach to structural estimation. *arXiv preprint arXiv:2007.06169*.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.