AIPW & NF

Chanhyuk Park Xiangyu Song

March 2025

1 Introduction

In observational studies, political scientists care about an unbiased average treatment effect (ATE) or average treatment effect on the treated (ATT). One way to generate an biased and also semiparametrically efficient estimator is by using augmented inverse propensity weight (AIPW), known to be doubly robust. However, there are two caveats here. First, as continuous treatment is common in political science studies, simply dichotomized treatments to fit into AIPW with binary treatment framework may lose information of the treatment effect heterogeneity. Second, although researchers can use logit or probit to estimate propensity score with binary treatment easily, an accurate estimation of propensity score under continuous treatment is still a challenge. In the paper, we propose to use the normalizing flows (NF), which produce efficient and exact propensity score estimation with continuous treatment.

2 Augmented Inverse Propensity Weighting

In observational studies, the key assumption is the unconfoundedness. Formally, let X_i denote a vector of pretreatment variables or covariates, D_i denote the treatment, and $Y_i(0), Y_i(1)$ denote the potential outcomes. We assume that the treatment is independent of the potential outcomes given X_i.

Assumption 1 (Unconfoundedness)**.**

$$D_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i$$

In applications, when the dimension of pretreatment variables X_i is substantial, conditioning on X_i can be a challenge. One key result from Rosenbaum and Rubin (1983) shows that Assumption 1 implies

$$D_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid e(X_i),$$

where $e(\cdot)$ is the propensity score that plays a key role in observational studies.

To identify the average causal effect under unconfoundedness, we also need to ensure that we can estimate the average effect at every value of the covariates, requiring overlap, that guarantees that the treatment assignment probability is bounded away from zero and one.

Assumption 2 (Overlap)**.**

$$0 < e(X_i) \equiv \Pr(D_i = d \mid X_i = x) < 1.$$

To estimate the causal effect of interest, e.g., ATE or ATT, a variety of methods have been proposed under the above unconfoundedness and overlap assumptions, referred as strong ignorability by Rosenbaum and Rubin (1983). Imbens and Xu (2024) summarise these methods into three groups.

The first one is outcome modeling. Define the conditional expectation of the two potential outcomes given treatment and covariates as

$$\mu_d(x) = \mathbb{E}[Y_i(d) \mid X_i = x], \quad \text{where} \quad d \in \{0, 1\}.$$

Then, under the unconfoundedness assumption, we can define the population ATE as

$$\tau^{\text{pop}} = \mathbb{E}[\mu_1(X) - \mu_0(X)].$$

The second group of methods focuses on adjusting covariate imbalance between the treatment and control groups. Among various blocking, matching, and reweighting methods, inverse propensity weighting (IPW) leverages the propensity score instead of simple covariate matching, which suffers from the curse of dimensionality. Rather than viewing the propensity score as a balancing score, exploiting the interpretation as the probability of being exposed to the treatment gives the following result when assuming unconfoundedness:

$$\mathbb{E}\left[\frac{DY}{e(X)}\right] = \mathbb{E}[Y(1)], \quad \mathbb{E}\left[\frac{(1-D)Y}{1-e(X)}\right] = \mathbb{E}[Y(0)].$$

Reweighting the units by the inverse of propensity score, we can get the sample analogy of the ATE:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)}\right],$$

where $\hat{e}(X_i)$ is the estimated propensity score. When the propensity is known, then this IPW estimator is unbiased for the ATE, and if propensity scores are estimated consistently, then this estimator is consistent for ATE. However, such a simple IPW estimator is believed to have poor small sample properties when the propensity score gets close to zero or one for some observations (Glynn and Quinn, 2010).

To improve this small sample performance, scholars have developed various mixed methods that combine outcome modeling with methods addressing covariate imbalance to eliminate remaining biases or improve precision. Abadie and Imbens (2011) provide the example that while the bias of a simple matching estimator might dominate variance in high-dimensional cases, adding regression to account for the remaining imbalance can substantially reduce such biases. The AIPW estimator for ATE is as the following:

$$\begin{aligned}\tau_{\text{AIPW}} &= \mu_1(X) - \mu_0(X) + \frac{(Y - \mu_1(X))D}{e(X)} - \frac{(Y - \mu_0(X)(1-D)}{1-e(X)} \\ &= \frac{DY}{e(X)} - \frac{(1-D)Y}{1-e(X)} + \frac{e(X) - D}{e(x)(1-e(x))}\left[\mu_1(X)(1-e(X)) + \mu_0(X)e(X)\right].\end{aligned} \tag{1}$$

We can get the sample analogy as

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)} + \frac{\hat{e}(X_i) - D_i}{\hat{e}(X_i)(1-\hat{e}(X_i))}\left[\mu_1(X_i)(1-\hat{e}(X_i)) + \mu_0(X_i)\hat{e}(X_i)\right]\right\}.$$

The AIPW estimators can be viewed as combining an outcome model with an adjustment term, which consists of an IPW estimator applied to the residuals from the outcome model, as shown in Equation 1. Robins and Ritov (1997) use the term "double robustness" for these mixed methods. If either the propensity score or regression model is correctly specified parametrically, the AIPW estimator that combines weighting and regression is consistent.

To get the unbiased AIPW estimator, researchers may need to estimate the propensity score. When the treatment is binary, fitting a logit or probit model can provide the estimated propensity score. However, when the treatment is multinomial or continuous, we need more tools for the estimation. Here our focus is the continuous treatment, which is quite common in the political science setting. Existing studies mostly use kernel density estimation (KDE) technique for the estimation of propensity score. However, KDE largely depends on the chosen kernel and bandwidth. In the paper, we are proposing to use the NF, which are generative models that produce tractable distributions where both sampling and density evaluation can be efficient and exact for propensity score estimation (Kobyzev et al., 2021).

3 Normalizing Flows

NF are generative models which produce tractable distributions where both sampling and density evaluation can be efficient and exact, using invertible functions. In other words, it is a technique that maps a random variable with a complex probability density function to the simpler density function.

A typical NF procedure starts with a simple base distribution such as Gaussian and then attempts to transform that function into a more complex distribution by applying a finite number of invertible functions[1]. The process utilized the *change of variables formula*, which guarantees mapping of one probability density function for a random variable Y with another probability density function of random variable Z:

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= p_{\mathbf{Z}}(\mathbf{f}(\mathbf{y}))|\det \mathrm{D}\,\mathbf{f}(\mathbf{y})| \\ &= p_{\mathbf{Z}}(\mathbf{f}(\mathbf{y}))|\det \mathrm{D}\,\mathbf{g}(\mathbf{f}(\mathbf{y}))|^{-1}, \end{aligned}$$

where $\mathbf{g}$ is an invertible function, $\mathbf{f} = \mathbf{g}^{-1}$, and $\mathrm{D}\,\mathbf{f}(\mathbf{y}) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$ is the Jacobian of $\mathbf{f}$. Figure 1 illustrates this relationship. Since $\mathbf{f}$ moves in the direction of a simple base distribution, it is called *normalizing flow*.

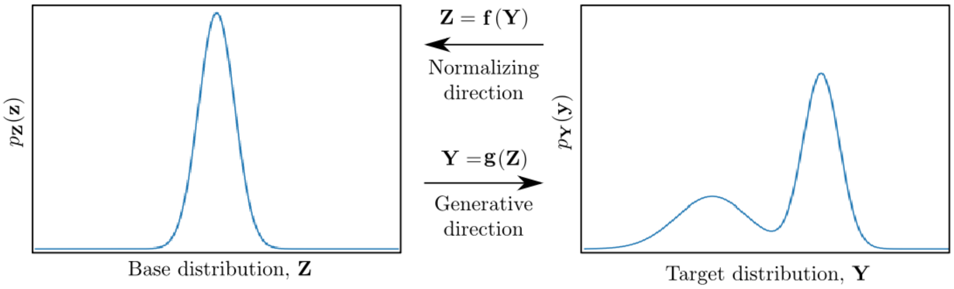
$p_{\mathbf{Z}}(\mathbf{z})$
Base distribution, $\mathbf{Z}$
$\mathbf{Z} = \mathbf{f}(\mathbf{Y})$
Normalizing direction
$\mathbf{Y} = \mathbf{g}(\mathbf{Z})$
Generative direction
$p_{\mathbf{Y}}(\mathbf{y})$
Target distribution, $\mathbf{Y}$

Figure 1: Graphical Illustration of Normalizing Flows. from Kobyzev et al. (2021)

[1] The existence of this process is always guaranteed, and discussed in Chen and Gopinath (2000).

As a result of this process, we can link the complex distribution (such as a joint distribution of variables we observe) with simple base distribution such as Gaussian. Since this is an invertible process, instead of dealing with the complex distribution itself, we can work with the simple base distribution such as sampling and density estimation, and then easily bring the result to the complex one.

NF is a useful tool in density estimation compare to other frequently used density estimation method. Histogram is a simple exploratory method. However, it relies heavily on the size of bins and thus often performs poorly on complex distributions. Unlike the histogram, KDE technique produces a smooth estimate of the density, using all sample points' locations. The main idea behind KDE has been proposed long ago (Rosenblatt, 1956), but the recent development of computing technology makes it a popular option today. KDE, however, requires researchers to decide which kernel to be used and the size of the bandwidth, and both influence the estimation result. Since NF usually relies on neural network, the computational cost may be high for complex design, it is more capable of dealing with complex density.

4 Using NF for Propensity Score Estimation

One straight forward application of density estimation with NF is propensity score estimation. Statistical estimators employing propensity score relies heavily on the accuracy of the propensity score.

Although there are criticism on using machine learning algorithm in propensity score estimation, it becomes more popular (Goller et al., 2020). Lechner and Okasa (2024) utillized random forest estimation of propensity score. In this paper, we propose to use NF as an alternative for propensity score estimation and leverage the estimated propensity score into the AIPW estimator in observational studies.

References

ABADIE, A. AND G. W. IMBENS (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11.

CHEN, S. AND R. GOPINATH (2000): "Gaussianization," in *Advances in Neural Information Processing Systems*, ed. by T. Leen, T. Dietterich, and V. Tresp, MIT Press, vol. 13.

GLYNN, A. N. AND K. M. QUINN (2010): "An Introduction to the Augmented Inverse Propensity Weighted Estimator," *Political Analysis*, 18, 36–56.

GOLLER, D., M. LECHNER, A. MOCZALL, AND J. WOLFF (2020): "Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed," *Labour economics*, 65, 101855.

IMBENS, G. W. AND Y. XU (2024): "LaLonde (1986) after Nearly Four Decades: Lessons Learned," *arXiv*.

KOBYZEV, I., S. J. PRINCE, AND M. A. BRUBAKER (2021): "Normalizing Flows: An Introduction and Review of Current Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 3964–3979.

LECHNER, M. AND G. OKASA (2024): "Random Forest estimation of the ordered choice model," *Empirical Economics*.

ROBINS, J. M. AND Y. RITOV (1997): "Toward a Curse of Dimensionality Appropriate (Coda) Asymptotic Theory for Semi-Parametric Models," *Statistics in Medicine*, 16, 285–319.

ROSENBAUM, P. R. AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*.

ROSENBLATT, M. (1956): "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, 27, 832–837.