Augmented Inverse Propensity Weighting and Normalizing Flows

Chanhyuk Park   Xiangyu Song

Washington University in St. Louis

Motivation and Research Question

In political science, researchers often seek unbiased estimates of the treatment effect. Augmented inverse propensity weighting (AIPW) offers an unbiased and semiparametrically efficient estimator. However, two challenges remain for continuous treatment: (1) dichotomizing continuous treatments to fit binary AIPW frameworks risks losing information on treatment effect heterogeneity, and (2) accurately estimating propensity scores with continuous treatments is difficult. To address these issues, we propose using normalizing flows (NF) for efficient and precise propensity score estimation under continuous treatment.

Problem Setup

We consider an outcome $Y \in \mathcal{Y} \in \mathbb{R}$, a univariate continuous treatment $T \in \mathcal{T} \subset \mathbb{R}$, a vector of pretreatment covariates $\mathbf{X} = (X_1, \dots, X_d) \in \mathcal{S} \subset R^d$. The Lebesgue density of the tuple $(Y, T, \mathbf{X})$ is $p(y, t, \mathbf{x}) = p_{Y|T,\mathbf{X}}(y \mid t, \mathbf{x}) \cdot p_{T|\mathbf{X}}(t \mid \mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x})$.

AIPW with Continuous Treatment

We focus on observational studies with continuous treatment. Suppose we have the data drawn i.i.d. from the joint distribution of $(Y, T, \mathbf{X})$ generated following

$$Y = \mu(T, \mathbf{X}) + \epsilon, \tag{1}$$

where $\epsilon \in \mathbb{R}$ with $\mathbb{E}[\epsilon \mid \mathbf{X}] = 0$ and $\mathrm{Var}(\epsilon) > 0$. We have potential outcome model

$$Y(t) = \mu(t, \mathbf{X}) + \epsilon,$$

for any treatment value t. Our main estimands of interest are

$$m(t) := \mathbb{E}[Y(t)] \quad \text{and} \quad \theta(t) := \frac{d}{dt}\mathbb{E}[Y(t)],$$

where $\mathbb{E}[Y(t)]$ is the average causal effect curve or dose-response curve, $\theta(t)$ is the derivative effect curve [2, 6].

Assumptions

1. (Consistency) $T = t$ implies $Y = Y(t)$ for any $t \in \mathcal{T}$.
2. (Unconfoundedness) $Y(t) \perp\!\!\!\perp T \mid \mathbf{X}$ for all $t \in \mathcal{T}$.
3. (Treatment Variation) The conditional variance of T given $\mathbf{X} = \mathbf{x}$ is strictly positive for all $\mathbf{x} \in \mathcal{X}$, i.e., $\mathrm{Var}(T \mid \mathbf{X} = \mathbf{x}) > 0$.
4. (Interchangeability) The equality $\frac{d}{dt}\mathbb{E}[\mu(t, \mathbf{X})] = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \mathbf{X})\right]$ holds true under model (1).
5. (Overlap) For $c > 0$, $0 < p_{T|\mathbf{X}}(t \mid \mathbf{x}) < 1 - c$, for all $(t, \mathbf{x}) \in \mathcal{T} \times \mathcal{X}$.

Identification

Under these assumptions and conditional expectation function (CEF) $\mu(t, \mathbf{x}) = \mathbb{E}[Y \mid T = t, \mathbf{X} = \mathbf{x}]$, we can identify $m(t)$ and $\theta(t)$ as

$$m(t) = \mathbb{E}[\mu(t, \mathbf{X})] \quad \text{and} \quad \theta(t) = \mathbb{E}\left[\frac{\partial}{\partial t}\mu(t, \mathbf{X})\right].$$

Following [6], we aim to obtain the following doubly robust AIPW estimators:

$$\hat{m}_{\mathrm{DR}}(t) = \frac{1}{nh}\sum_{i=1}^{n}\left\{\frac{K\left(\frac{T_i - t}{h}\right)}{\hat{p}_{T|\mathbf{X}}(T_i \mid \mathbf{X}_i)} \cdot [Y_i - \hat{\mu}(t, \mathbf{X}_i)] + h \cdot \hat{\mu}(t, \mathbf{X}_i)\right\} \tag{2}$$

$$\hat{\theta}_{\mathrm{DR}}(t) = \frac{1}{nh}\sum_{i=1}^{n}\left\{\frac{\left(\frac{T_i - t}{h}\right) K\left(\frac{T_i - t}{h}\right)}{h \cdot \kappa_2 \cdot \hat{p}_{T|\mathbf{X}}(T_i \mid \mathbf{X}_i)}\left[Y_i - \hat{\mu}(t, \mathbf{X}_i) - (T_i - t) \cdot \hat{\beta}(t, \mathbf{X}_i)\right] + h \cdot \hat{\beta}(t, \mathbf{X}_i)\right\}, \tag{3}$$

where $\hat{\beta}(t, \mathbf{X}_i)$ is a consistent estimator of $\beta(t, \mathbf{x}) = \frac{\partial}{\partial t}\mu(t, \mathbf{x})$, $K : \mathbb{R} \to [0, \infty)$ is a kernel function with $\kappa_2 = \int u^2 K(u)du$, $h > 0$ is a smoothing bandwidth, and $\hat{p}_{T|\mathbf{X}}(t \mid \mathbf{x})$ is a consistent estimator of $p_{T|\mathbf{X}}(t \mid \mathbf{x})$.

Our goal is to estimate $\hat{p}_{T|\mathbf{X}}(t \mid \mathbf{x})$ using NF instead of kernel density estimation (KDE) which is commonly used in the literature.

Normalizing Flows

- Generative models: produce tractable distributions using invertible functions.
- Both sampling and density evaluation can be efficient and exact,
- Maps a random variable with a complex probability density function to the simpler density function.

1. Starts with a **simple base distribution** such as Gaussian
2. Transform that function into a more complex distribution by applying a finite number of **invertible functions**.

The process utilized the *change of variables formula*, which guarantees mapping of one probability density function for a random variable $\mathbf{Y}$ with another probability density function of random variable $\mathbf{Z}$:

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

where $\mathbf{g}$ is an invertible function, $\mathbf{f} = \mathbf{g}^{-1}$, and $\mathbf{f}(\mathbf{y}) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$ is the Jacobian of $\mathbf{f}$.

Simulation for NF

- Tests of uncovering conditional distribution $p_{Y|\mathbf{X}}(y \mid \mathbf{x})$
  - Y: Gaussian Mixture, Ring Mixture and Two Moons
  - X: 5 dimensional vector (Gaussian distributions)
- Using a sample of 1,000 data points from the target distribution
- Flow: Masked Autorsegressive Normalizing Flow (MANF) with 4 layers, 128 hidden units [3]
- The Base Distribution: Resampled Gaussian distribution based on Latent Accept Reject Sampling (LARS)[1, 5] with 2 hidden layer (256 units each) MLP model with ReLU and Sigmoid activation
- Trained with bootstrap resampling boosting
- The loss is measured in forward KL divergence [4]

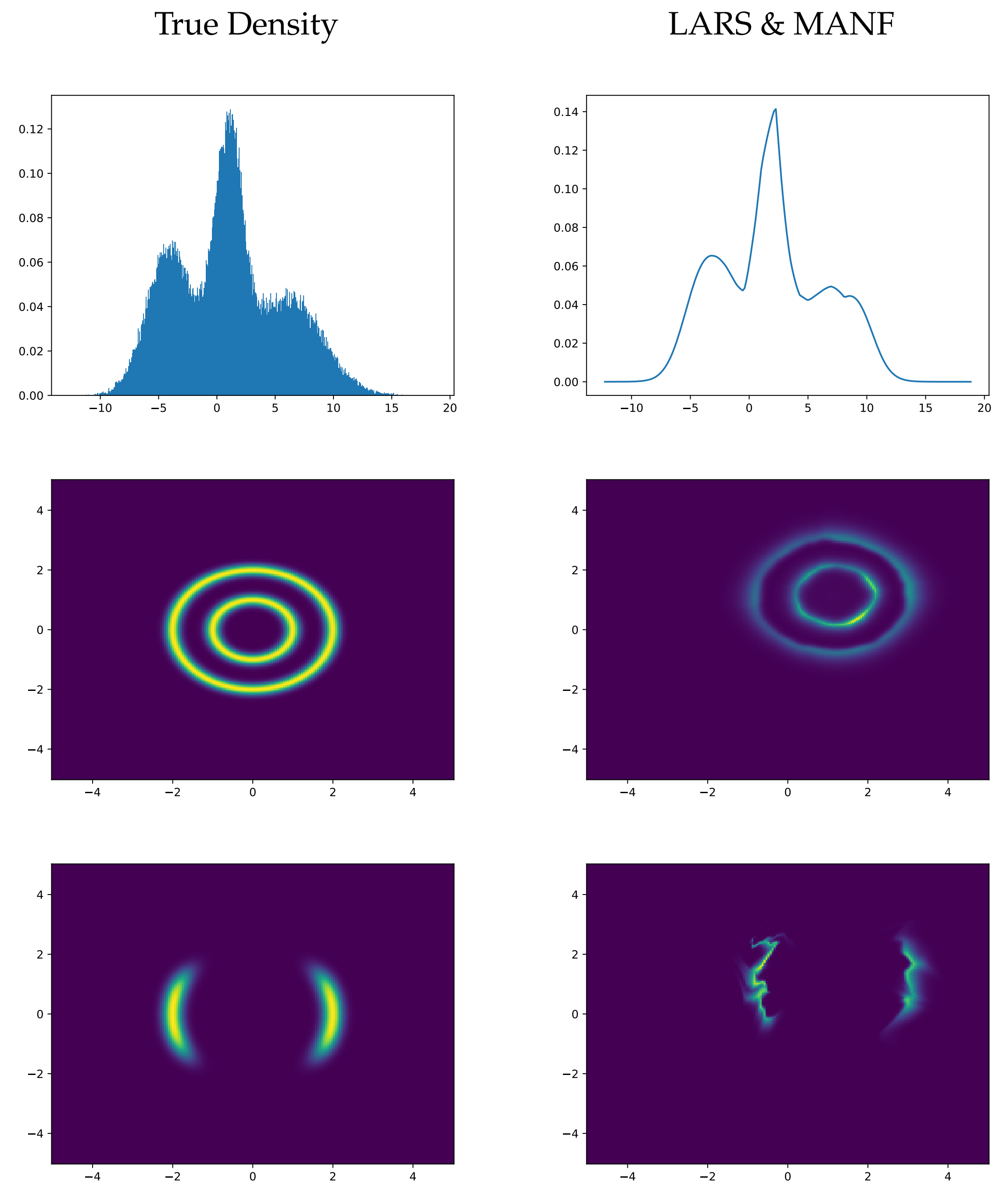

Figure 1. 1D and 2D Toy Example of LARS Normalizing Flow

Simulation for AIPW with NF

We generated i.i.d observations $\{(Y_i, T_i, \mathbf{X}_i)\}_{i=1}^{n}$ from the following DGP:

$$\begin{aligned} &Y = 1.2T + T^2 + TX_1 + 1.2\boldsymbol{\xi}^\top \mathbf{X} + \epsilon \cdot \sqrt{0.5 + \Phi(X_1)}, \quad \epsilon \sim \mathcal{N}(0, 1) \\ &T = \Phi(3\boldsymbol{\xi}^\top \mathbf{X}) - 0.5 + 0.75E, \quad \mathbf{S} = (S_1, \dots, S_d)^\top \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}), \quad E \sim \mathcal{N}(0, 1), \end{aligned} \tag{4}$$

where Φ is the CDF of $\mathcal{N}(0, 1)$, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d) \in \mathbb{R}^d$ has entry $\xi_j = \frac{1}{j^2}$ for $j = 1, \dots, d$ and $\Sigma_{ii} = 1, \Sigma_{ij} = 0.5$ for $|i - j| = 1$ and $\Sigma_{ij} = 0$ for $|i - j| > 1$. We set $d = 20$. Our estimands of interests are given by $m(t) = 1.2t + t^2$ and $\theta(t) = 1.2 + 2t$.

- Flow: Masked Autoregressive Flow with 32 layers and 64 hidden units
- Base Distribution: Resampled Gaussian with LARS (2 hidden layers of size 256)
- Training: Bootstrap boosting with sample sizes of 500, 1,000, and 2,000.

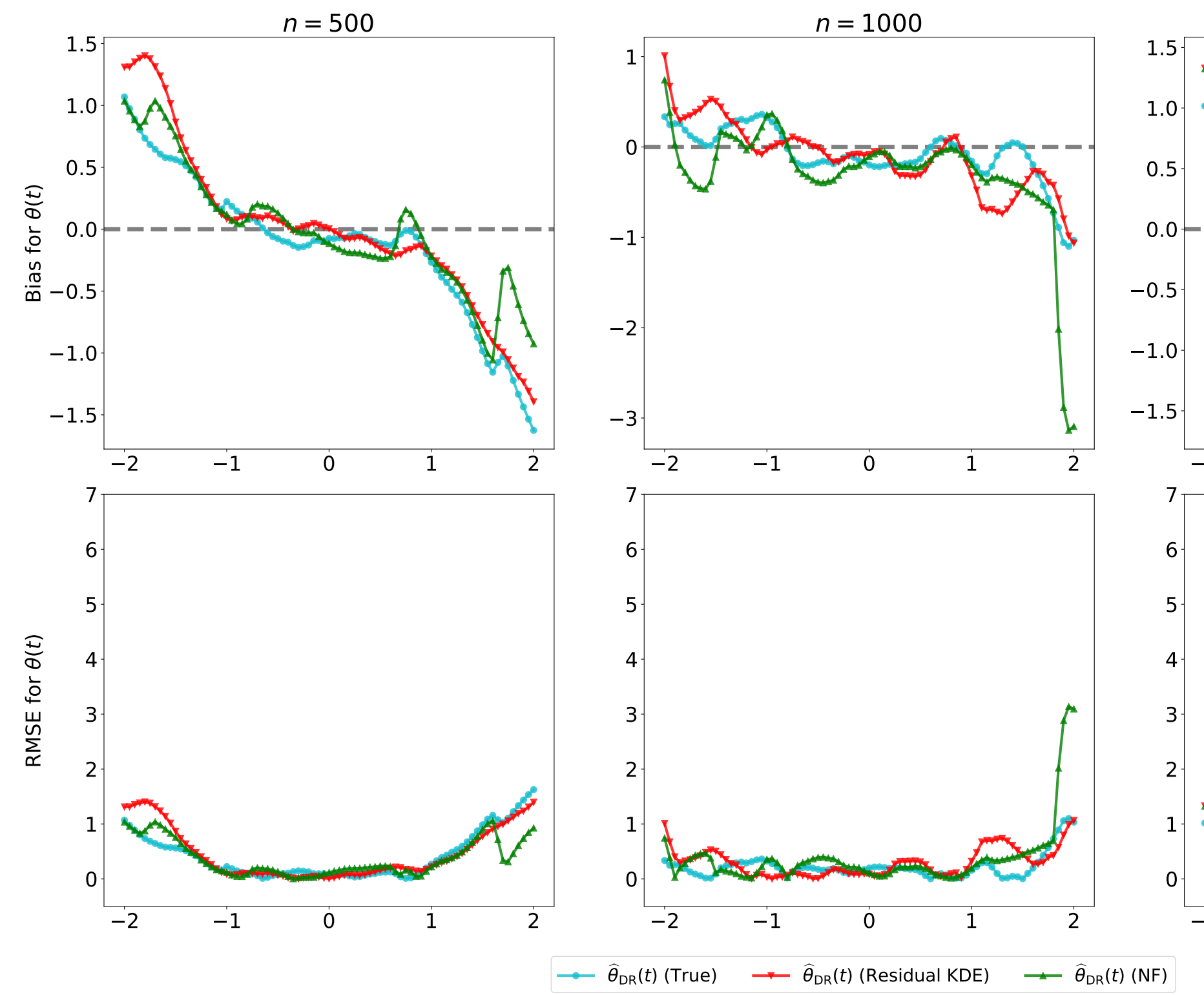

Figure 2. $\theta(t)$: Comparison between Residual KDE and Our Proposed Method

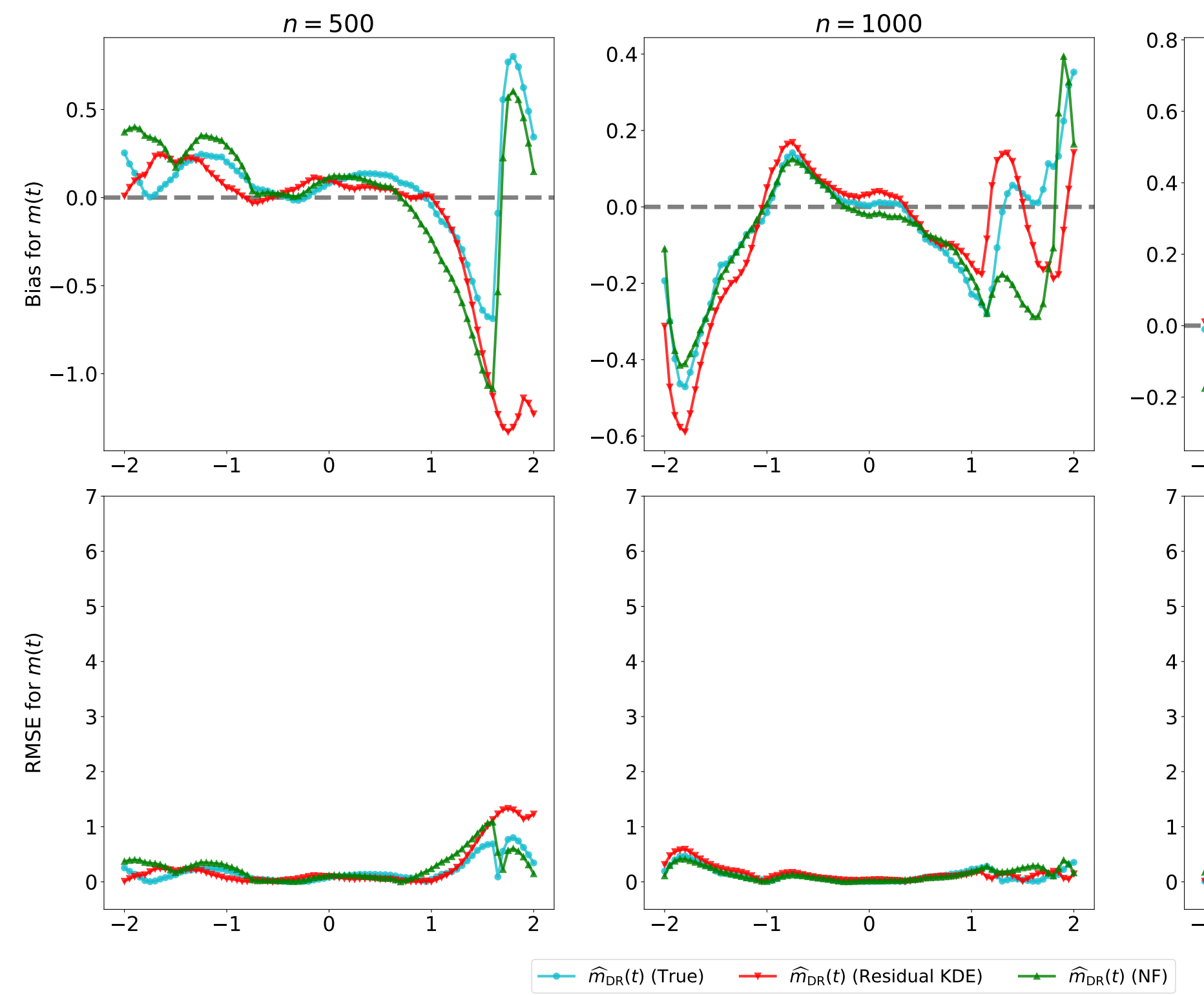

Figure 3. $m(t)$: Comparison between Residual KDE and Our Proposed Method

Conclusion

- Extend AIPW with continuous treatment proposed by [6].
- Unlike residual KDE, our NF works without assuming additive errors.
- Hyperparameter tuning and a more sophisticated ensemble technique may help enhance performance.

References

For more results and codes, visit: `https://github.com/chanhyuk58/aipw_nf`.

[1] M. Bauer and A. Mnih. Resampled priors for variational autoencoders. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 66–75. PMLR, 16–18 Apr 2019.

[2] GW. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. doi: 10.1093/biomet/87.3.706.

[3] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[4] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2021.

[5] V. Stimper, B. Schölkopf, and J. Miguel Hernandez-Lobato. Resampling base distributions of normalizing flows. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4915–4936. PMLR, 28–30 Mar 2022.

[6] Y. Zhang and Y.-C. Chen. Doubly Robust Inference on Causal Derivative Effects for Continuous Treatments. *arXiv preprint arXiv:2501.06969*, 2025. doi: 10.48550/arXiv.2501.06969.

chanhyuk@wustl.edu,   xiangyusong@wustl.edu