# Evaluating the Robustness of INFs Under Nuisance Misspecification

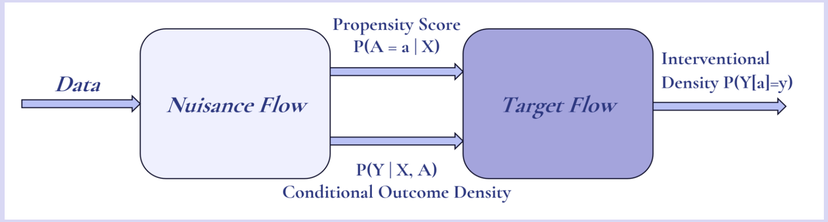
Ruthvik Allu - rallu@student.tudelft.nl

Supervisors: Jesse Krijthe, Rickard Karlsson

Faculty: EEMCS, Delft University of Technology, The Netherlands

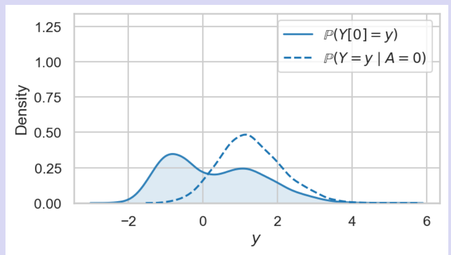
## Introduction

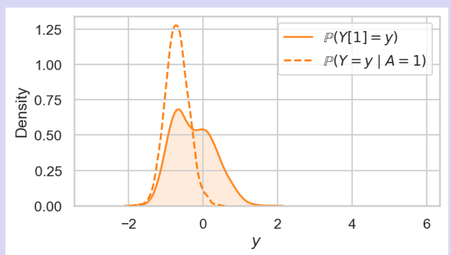
In real-world decision-making, understanding causal effects - not just correlations - is essential. Interventional Normalizing Flows (INFs) are a recent deep learning method that estimates full interventional outcome distributions from observational data. INFs consist of two components: the nuisance flow, which models the propensity score and conditional outcome distribution, and the target flow, which uses these to estimate interventional densities. INFs are theoretically doubly robust, meaning accurate estimates can still be achieved if either the propensity score or outcome model is correctly specified. This research investigates the practical robustness of INFs by systematically testing how errors in the nuisance flow impact the quality of interventional estimates.




## Methods

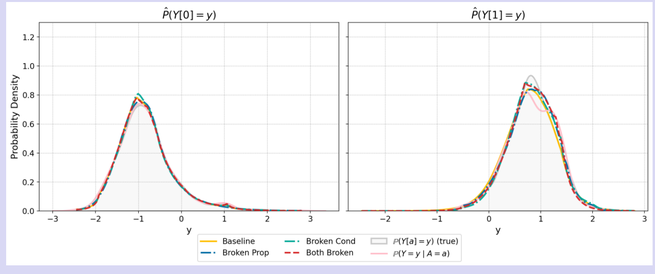
This study evaluates the robustness of INFs using four benchmark datasets: *GaussianClean* and *BimodalClean*, which have low confounding and relatively simple to moderately complex outcome distributions; *SyntheticComplex and SplitPeaks*, which introduce stronger confounding. Across all experiments, the official INF implementation and author-recommended hyperparameters are used as the baseline. The robustness of INFs is assessed in three ways: (1) first, by completely breaking components of the nuisance flow - either the propensity score model, the conditional outcome model, or both - to evaluate extreme failure cases; (2) second, by perturbing key hyperparameters of the nuisance flow to simulate more minor misspecifications; and (3) third, by injecting controlled noise into the nuisance flow outputs to test sensitivity to imperfect estimates. All results are compared to the baseline INF estimates to measure the degradation in interventional distribution quality.

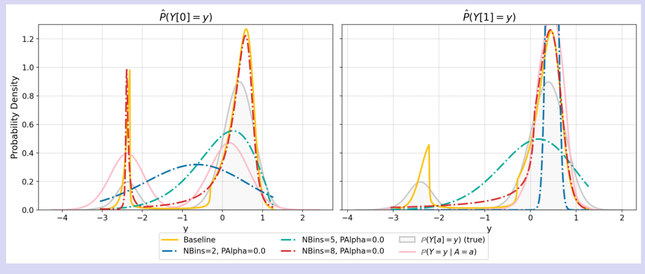



## Results

In low-confounding datasets (GaussianClean and BimodalClean), INF estimates remained accurate even when the nuisance flow components were completely broken or mildly perturbed. However, in high-confounding settings (SplitPeaks and SyntheticComplex), breaking either the propensity score or conditional outcome model led to significant deviations from the baseline, with the conditional outcome model having a greater impact. Hyperparameter perturbations showed that extreme values, especially low expressiveness (few bins) and absence of the propensity loss, could severely degrade performance. Noise injection experiments revealed that moderate to high noise notably affected estimates in high-confounding scenarios, while low noise had little effect across datasets.




## Discussion

The results show that the robustness of INFs depends heavily on the level of confounding in the data. In low-confounding datasets, the model maintains reliable estimates even when the nuisance flow is heavily misspecified. However, in some high-confounding settings, the conditional outcome model becomes especially critical. Breaking it leads to substantial degradation in estimate quality, more than breaking the propensity score. The doubly robust property appears to hold in ideal or mildly misspecified cases but weakens under stronger bias or when both nuisance components are compromised. Additionally, while INF handles small perturbations and low levels of noise well, larger deviations - either in hyperparameters or noise - can lead to unstable results.

## Conclusion

This study provides a focused empirical evaluation of INF robustness under nuisance misspecification. The findings suggest that while INFs are resilient in low-bias settings, their reliability diminishes in more complex, confounded scenarios. Although minor modeling imperfections are generally tolerated, careful validation of the nuisance flow is critical for trustworthy causal estimates in real-world applications.