Comparison of different Latent Source Distributions in Causal Graphical Normalizing Flows on the Estimated Causal Effect

Proposal for a Master Thesis in Statistics and Machine Learning at Linköping Univeristy, Sweden

1 Background

Answering causal questions is fundamental in many different fields of science like medicine or social sciences. Answering these questions requires researchers to perform randomized control trials, which is considered to be the gold standard for causal effect analyses. However, performing experiments might not always be possible e.g. due to cost or ethical reasons, which is when using observational data to answer causal questions becomes relevant.

To be able to estimate the causal effect of a treatment variable T on a outcome variable Y from observational data requires having access to observations of all confounders between T and Y, i.e. it is assumed that there are no *hidden* confounders. Models like causal normalizing flows (cGNFs) can be used to estimate the data generating process [2] by learning a transformation that maps variables from any given latent distribution of the source space ε to the observed space (T, Y). The setup can be seen in Figure 1. In this setting, the distribution of the latent space is representative of different functional forms of how confounders impact the variables T and Y.

Sensitivity analysis is a field of causal inference where the violation of the assumption that there are no *hidden* confounders is investigated [4]. Previous research has proposed to assess the impact of this violation by assuming a Gaussian distribution of the latent space and estimating the causal effect under different values of the correlation coefficient of the latent Guassian [1]. The proposed thesis aims at assessing the impact of using other latent distributions in a cGNF on the data generating process it produces.

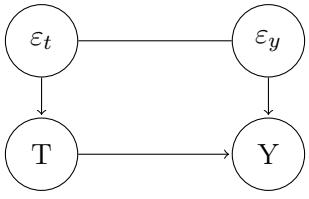

Figure 1: Causal Graph of the Causal Estimation Problem

2 Research Questions

1. How can the training process of a cGNF be adapted for different latent distribuions of ε?

2. What are suitable synthetic datasets to compare different assumed latent distributions of ε?

3. How big is the difference in causal effect estimations (i.e. in the functional relationship between T and Y) between different assumed latent distributions for synthetic and real world datasets (e.g. [3])?

3 Datasets

One of the goals of the thesis is to define suitable synthetic datasets to compare the effect of different assumed latent distributions. As a real world dataset, the dataset by Blau and Duncan [3] is suggested. Blau and Duncan conducted a study about social mobility in the US. It contains information about individual's educational attainment (T) and occupational status (Y), where possible hidden confounders between the two variables are unobserved. cGNFs can be used to assess the causal effect of T on Y under different assumed latent distributions.

4 Eligibility Criteria

Passed the course 732A99 Machine Learning and it can be helpful to have passed 732A96 Advanced Machine Learning, even though not being a formal requirement.

5 Contact Person

If you are interested in this thesis, please contact Marc Braun, marc.braun@liu.se. If you have your own suggestions for theses on causal inference or conditional generative modelling, feel free to reach out to me too.

References

[1] Sourabh Balgi, Jose M. Peña, and Adel Daoud. *ρ-GNF: A Copula-based Sensitivity Analysis to Unobserved Confounding Using Normalizing Flows*. 2024. arXiv: 2209.07111 [stat.ME]. URL: https://arxiv.org/abs/2209.07111.

[2] Sourabh Balgi et al. *Deep Learning With DAGs*. 2024. arXiv: 2401.06864 [stat.ML]. URL: https://arxiv.org/abs/2401.06864.

[3] Peter M Blau, Otis Dudley Duncan, and A Tyree. *The American Occupational Structure. With the Collaboration of A. Tyree*. Wiley, 1967.

[4] Cheng Lin, Jose M. Pena, and Adel Daoud. *Assessing the Unobserved: Enhancing Causal Inference in Sociology with Sensitivity Analysis*. 2024. arXiv: 2311.13410 [stat.ME]. URL: https://arxiv.org/abs/2311.13410.