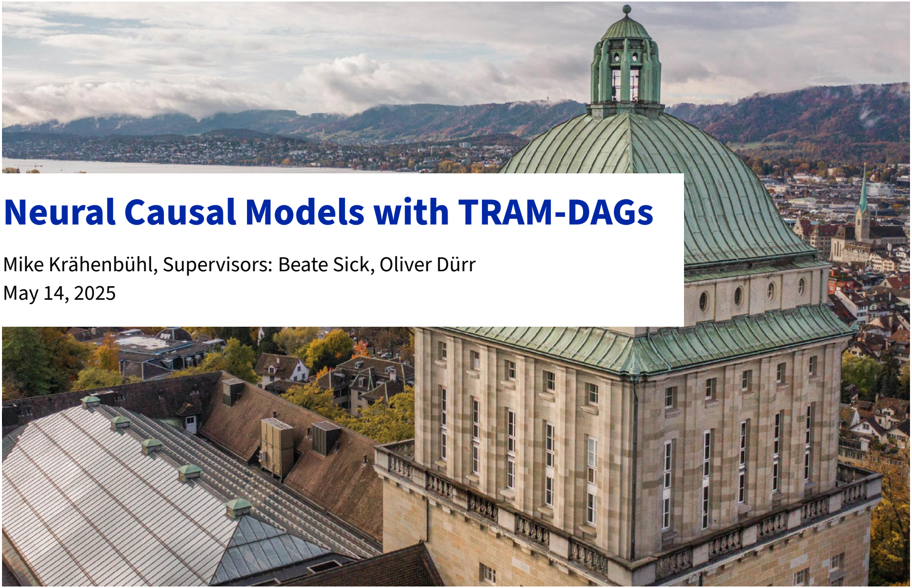
Neural Causal Models with TRAM-DAGs

Mike Krähenbühl, Supervisors: Beate Sick, Oliver Dürr

May 14, 2025



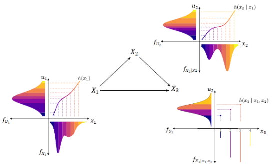
Background

Supervisors:

- Beate Sick, UZH
- Oliver Dürr, HTWG Konstanz

Paper *"Interpretable Neural Causal Models with TRAM-DAGs"* (Sick and Dürr, 2025):

- Framework to model causal relationships
- Based on transformation models
- Rely on (deep) neural networks
- Compromise between interpretability and flexibility



Research Questions

Sick and Dürr (2025) showed on synthetic data, that TRAM-DAGs can be fitted on observational data and tackle causal queries on all three levels of Pearl's causal hierarchy.

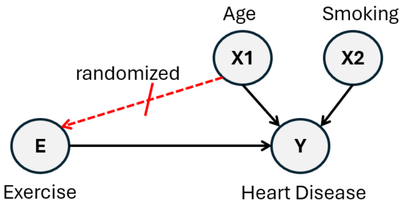
In my thesis:

1. Apply the framework on real-world data
   - DAG has to be defined
   - Ground-truth is not known
2. Individualized Treatment Effect estimation
   - Potential outcomes under different treatments
   - Crucial for personalized medicine

RCT vs. Observational Data

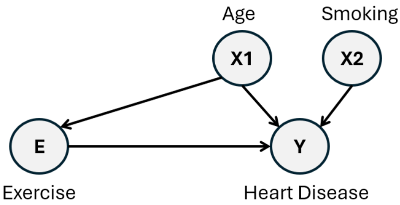
Randomized Controlled Trial:

- Gold standard for estimating causal effect
- Solves problem of confounding



Observational Data:

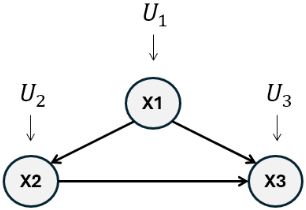
- Real world, potential confounding
- We assume no unobserved confounding



Structural Causal Model

SCM: Describes the causal mechanism and probabilistic uncertainty

- X_i = observed variable
- U_i = noise distribution



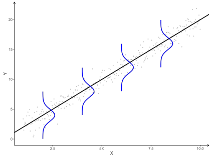
$$U_1 \sim F_{U_1}, U_2 \sim F_{U_2}, U_3 \sim F_{U_3}$$

$$X_1 = f_1(U_1)$$
$$X_2 = f_2(U_2, X_1)$$
$$X_3 = f_3(U_3, X_1, X_2)$$

Estimating Functional Form

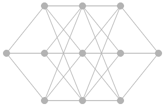
Statistical methods:

- E.g. linear/logistic regression
- Predefined form, risk of bias if misspecified



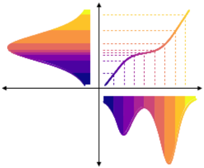
Neural networks:

- E.g. feed-forward NNs, normalizing flows, VACAs
- Flexible, but "black-box", data-type limitations



TRAM-DAGs:

- Compromise: flexibility + interpretability
- Mixed data types

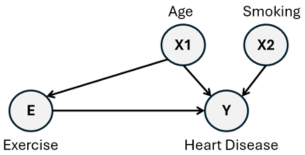


Pearl's Causality Ladder

Observational (seeing)

$P(Y = 1 \mid E = 1)$

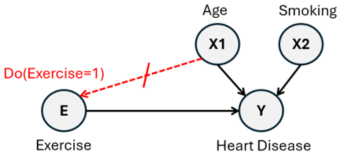
"Probability of heart disease given that the person exercises"



Interventional (doing)

$P(Y = 1 \mid \mathrm{do}(E = 1))$

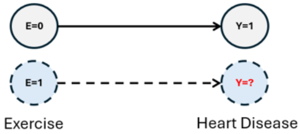
"Probability of heart disease if we made people start exercising"



Counterfactual (imagining)

$P(Y_{(E=1)} = 1 \mid E = 0, Y = 1)$

"Would someone who does not exercise and has heart disease still have it if they had exercised?"



Individualized Treatment Effect (ITE)

Difference in outcomes between two treatment options, for one specific individual with unique characteristics.

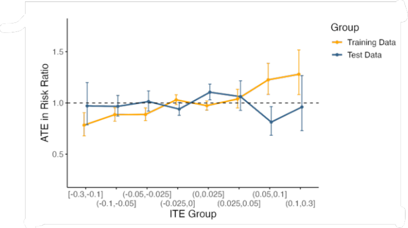
$$\mathrm{ITE}_i = P(Y_i = 1 \mid T = 1, \mathbf{X} = \mathbf{x_i}) - P(Y_i = 1 \mid T = 0, \mathbf{X} = \mathbf{x_i})$$

Difficulty:

- We can only observe one facutal outcome - the other one is counterfactual

Recent findings:

- Chen et al. (2025) analyzed mainstream causal ML methods for ITE estimation on two large RCTs.
- ITEs estimated from training data failed to generalize to the test data



Transformation Models

Flexible distributional regression method (Hothorn et al., 2014)

Continuous $Y \in \mathbb{R}$:

$$F_{Y|\mathbf{X}=\mathbf{x}}(y) = F_Z(h(y) + \mathbf{x}^\top \boldsymbol{\beta})$$

Discrete $Y \in \{y_1, y_2, \dots, y_K\}$:

$$P(Y \leq y_k \mid \mathbf{X} = \mathbf{x}) = F_Z(\vartheta_k + \mathbf{x}^\top \boldsymbol{\beta}), \quad k = 1, 2, \dots, K-1$$

- F_Z: CDF of the standard logistic distribution
- h: Transformation function, monotonically increasing
- $\mathbf{x}$: Predictors

Transformation Models

Continuous Y:

Intercept: Bernstein polynomial

$h_I(y) = \frac{1}{M+1} \sum_{k=0}^{M} \vartheta_k \, \mathrm{B}_{k,M}(y)$

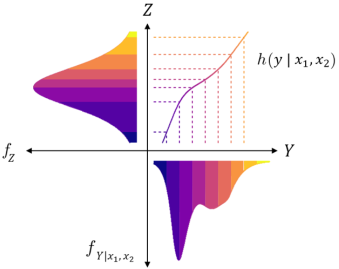
$h(y \mid \mathbf{x}) = h_I(y) - \mathbf{x}^\top \boldsymbol{\beta}$

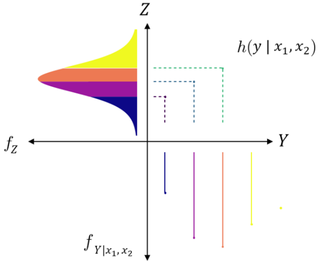
Discrete/Ordinal Y:

Intercept: Cut-off value

$h_I(y_k) = \vartheta_k$

$h(y_k \mid \mathbf{x}) = h_I(y_k) - \mathbf{x}^\top \boldsymbol{\beta}$

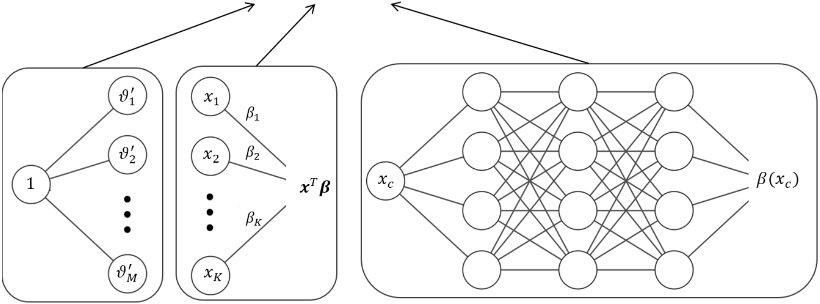




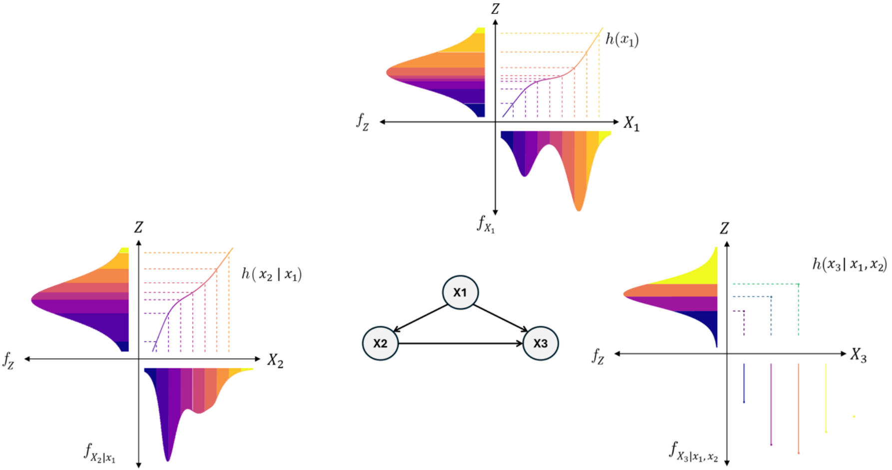
Deep TRAMs

- Extended to Deep TRAMs (Sick et al., 2021)
- Flexible components
- Minimize the NLL through NN optimization

$$h(y|\boldsymbol{x_L}, \boldsymbol{x_C}) = \mathrm{SI} + \mathrm{LS}_{\boldsymbol{x_L}} + \mathrm{CS}_{\boldsymbol{x_C}}$$



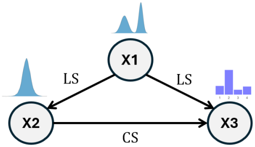
TRAM-DAGs



Z
$h(x_1)$
f_Z
X_1
f_{X_1}
Z
$h(x_2 \mid x_1)$
f_Z
X_2
$f_{X_2|x_1}$
X1
X2
X3
Z
$h(x_3 \mid x_1, x_2)$
f_Z
X_3
$f_{X_3|x_1,x_2}$

Simulation Example

- We have:
  - Observational data (simulated)
  - Predefined DAG
- We want:
  - Estimate conditional CDF of each variable
  - Sample from conditional distributions for causal queries

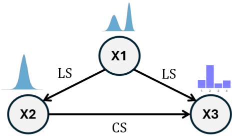


$$X_1 \sim F_z(h(x_1))$$
$$X_2 \sim F_z(h(x_2) + \mathrm{LS}_{x1})$$
$$X_3 \sim F_z(h(x_3) + \mathrm{LS}_{x1} + \mathrm{CS}_{x2})$$

Adjacency Matrix

Model structure represented by a meta-adjacency matrix:

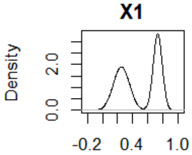
- **Rows**: source of effect
- **Columns**: target of effect



$$\longrightarrow \mathbf{MA} = \begin{bmatrix} 0 & \text{LS} & \text{LS} \\ 0 & 0 & \text{CS} \\ 0 & 0 & 0 \end{bmatrix}$$

Data Generating Process (DGP)

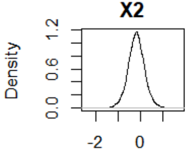
X_1: Continuous, bimodal. *Source node* (independent).



X_2: Continuous. Depends on X_1 (linear):

$$\beta_{12} = 2, \quad h_I(X_2) = 5X_2$$

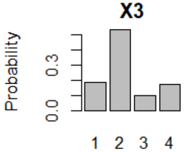
$$\boxed{h(X_2 \mid X_1) = h_I(X_2) + \beta_{12} X_1}$$



X_3: Ordinal. Depends on X_1 (linear) and X_2 (complex):

$$\beta_{13} = 0.2, \quad f(X_2) = 0.5 \cdot \exp(X_2), \quad \vartheta_k \in \{-2,\ 0.42,\ 1.02\}$$

$$\boxed{h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13} X_1 + f(X_2)}$$



Construct Model: Modular Neural Network

Inputs:

Observations + adjacency matrix

Outputs:

- Simple Intercepts (SI): ϑ
- Linear Shifts (LS): $\beta_{12}X_1$, $\beta_{13}X_2$
- Complex Shift (CS): $\beta(X_2)$

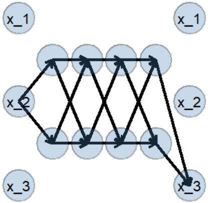
Transformation Functions:

$$\boxed{h(X_i \mid pa(X_i)) = \text{SI} + \text{LS} + \text{CS}}$$

$$h(X_1) = h_I(X_1)$$

$$h(X_2 \mid X_1) = h_I(X_2) + \beta_{12}X_1$$

$$h(X_{3,k} \mid X_1, X_2) = \vartheta_k + \beta_{13}X_1 + \beta(X_2)$$



CS_{X_2} on X_3

Loss: Negative Log-Likelihood (NLL)

CDF, density and NLL of the TRAM (for continuous outcome):

$$F_{Y|\mathbf{X}=\mathbf{x}}(y) = F_Z(h(s(y) \mid \mathbf{x}))$$

$$f_{Y|\mathbf{X}=\mathbf{x}}(y) = f_Z(h(s(y) \mid \mathbf{x})) \cdot h'(s(y) \mid \mathbf{x}) \cdot s'(y)$$

$$\text{NLL} = -\log(f_{Y|\mathbf{X}=\mathbf{x}}(y))$$

Standard logistic density:

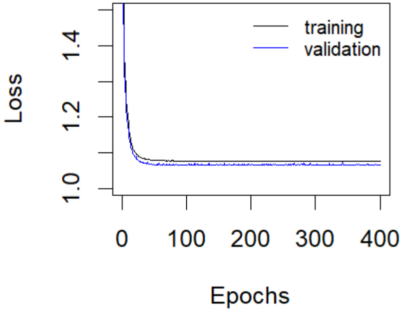
$$f_Z(z) = \frac{e^z}{(1+e^z)^2}, \quad z \in \mathbb{R}$$

Scaled y (Bernstein polynomial bounded $[0,1]$):

$$s(y) = \frac{y - \min(y)}{\max(y) - \min(y)}$$

Model Fitting

- Samples: 20’000 training / 5’000 validation
- Learning rate: 0.005
- Epochs: 400

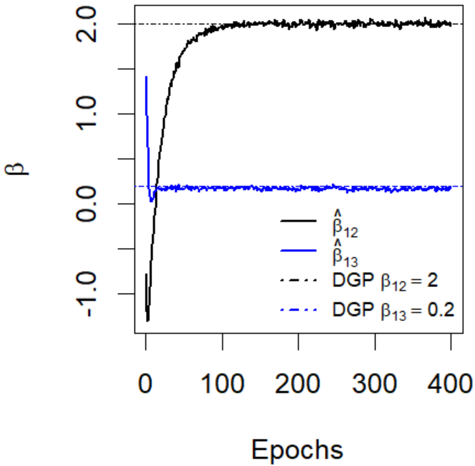


Interpretable Coefficients

Linear Shift Coefficients:

$$F(x_2 \mid x_1) = F_Z\left(h(x_2) - x_1\beta_{12}\right)$$

$$F(x_3 \mid x_1, x_2) = F_Z\left(h(x_3) - x_1\beta_{13} - CS_{x_2}\right)$$



Interpretable Coefficients

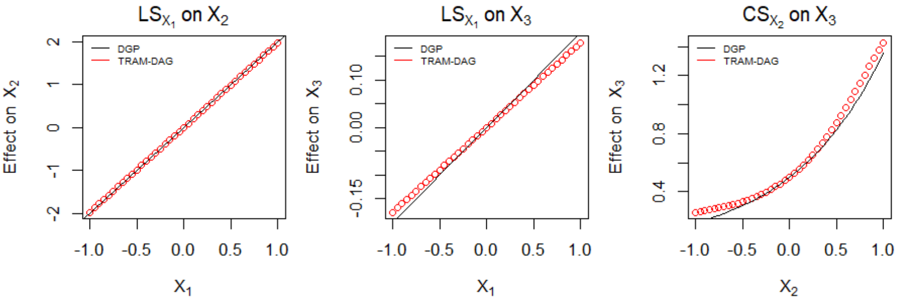
$$F_{X_2|X_1}(x_2) = \text{expit}(h(x_2) + \beta_{12}x_1)$$

$$\log\left(\frac{F_{X_2|X_1}(x_2)}{1 - F_{X_2|x_1}(x_2)}\right) = \text{expit}^{-1}(\text{expit}(h(x_2) + \beta_{12}x_1)) = h(x_2) + \beta_{12}x_1$$

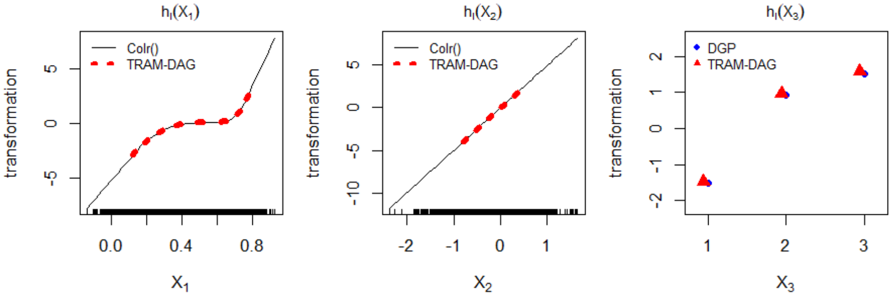
$$\text{OR}_{x_1 \to x_1+1} = \frac{\text{odds}(X_2 \leq x_2 \mid X_1 = x_1 + 1)}{\text{odds}(X_2 \leq x_2 \mid X_1 = x_1)} = \frac{\exp(h(x_2) + \beta_{12}(x_1 + 1))}{\exp(h(x_2) + \beta_{12}x_1)} = \exp(\beta_{12})$$

Interpretation:
$\exp(\beta_{12})$ is the **multiplicative change in odds** for $X_2 \leq x_2$ when increasing $\mathbf{X}_1$ by 1 unit, *holding all else constant*.

Linear and Complex Shifts



Intercepts



$h_I(X_1)$
Colr()
TRAM-DAG
transformation
5
0
-5
0.0
0.4
0.8
X_1
$h_I(X_2)$
Colr()
TRAM-DAG
transformation
5
0
-5
-10
-2
-1
0
1
X_2
$h_I(X_3)$
DGP
TRAM-DAG
transformation
2
1
0
-1
-2
1
2
3
X_3

What is Possible with a Fitted TRAM-DAG?

Once the model is fitted, it can be used:

- as genearative sampling model to sample observational and interventional distributions
- or to determine counterfactual outcomes in the continuous case.

Even if the model is fitted on observational data, we can make interventional and counterfactual statements, given the DAG is correct and no unobserved confounding.

Sampling from the Fitted TRAM-DAG (observational)

Nodes $X_i, i \in \{1,\ 2,\ 3\}$**:**

- Sample latent value:

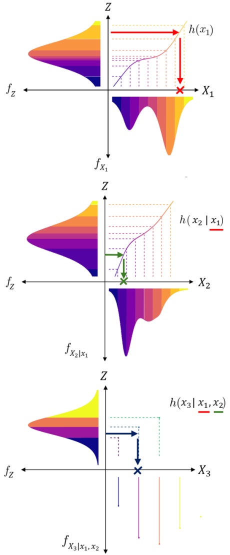
$$z_i \sim F_{Z_i} \quad \text{(e.g., } \texttt{rlogis()} \text{ in R)}$$

- Determine x_i such that:
  - **If X_i is continuous:**

$$x_i = h^{-1}(z_i \mid \mathrm{pa}(x_i))$$

  - **If X_i is ordinal:** find the smallest category x_i such that

$$x_i = \min\{x : z_i \leq h(x \mid \mathrm{pa}(x_i))\}$$

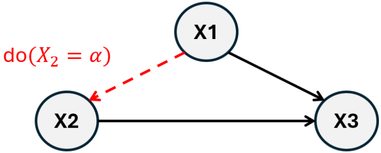


Sampling from the Fitted TRAM-DAG (interventional)

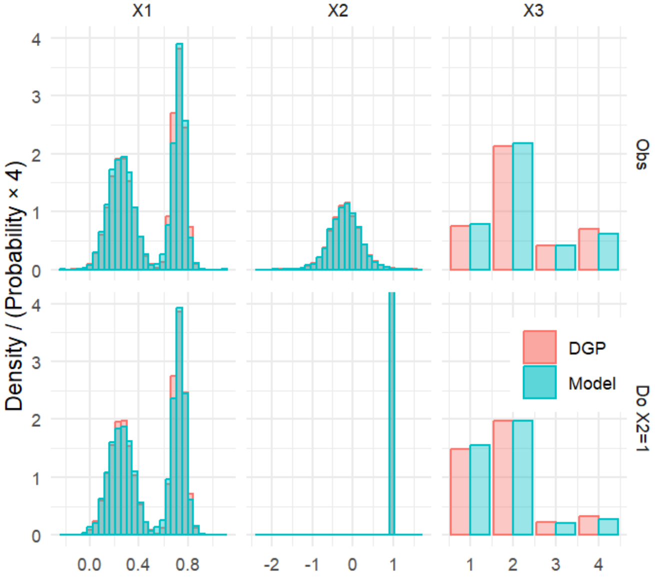
Interventional sampling:

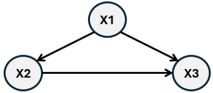
- Do-intervention: $\text{do}(x_2 = \alpha)$
- Sample from the interventional-distribution:

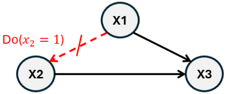
$$x_3 = \min\{x : z_3 \leq h(x \mid x_1, x_2 = \alpha)\}$$



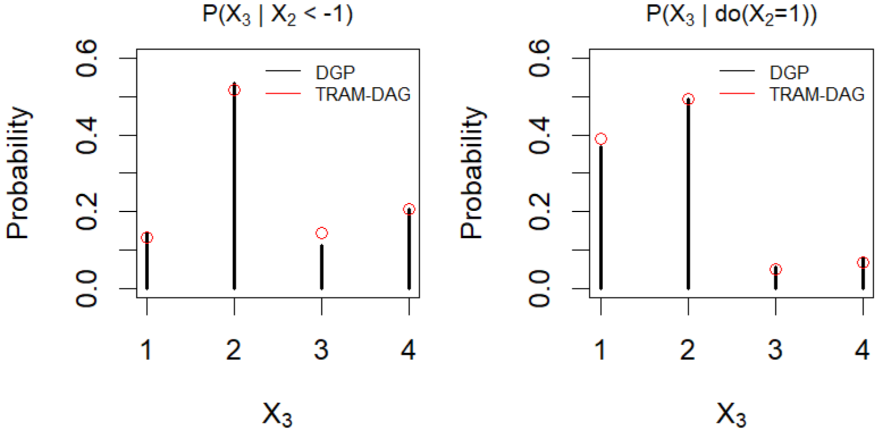
Sampling Distributions



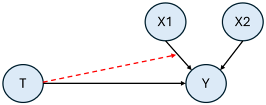




Observational and Interventional Queries

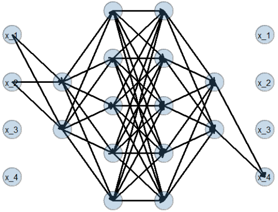


Example: ITE Estimation

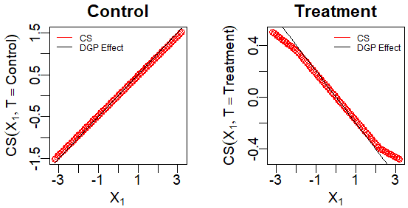


DGP: $\operatorname{logit}(P(Y=1 \mid T, X_1, X_2)) = \beta_0 + X_1\beta_1 + X_2\beta_2 + T\beta_3 + \color{red}{TX_1\beta_4}$

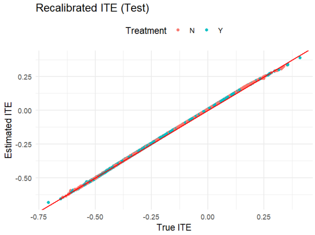
TRAM-DAG: $h(Y_k \mid T, X_1, X_2) = \vartheta_k + \color{red}{\mathrm{CS}(T, X_1)} + \mathrm{LS}(X_2)$

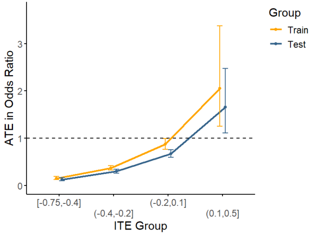


$\mathrm{CS}_{(T,X_1)}$ on Y



Example: ITE Estimation





ITE_i estimation from fitted model:

- $P(Y_i = 1 \mid \mathrm{do}(T=1), \mathbf{x_i})$
- $P(Y_i = 1 \mid \mathrm{do}(T=0), \mathbf{x_i})$
- Calculate ITE_i

Outlook

What we also did:

- Ordinal predictors
- TRAM-DAG on real climate data

What comes next:

- ITE estimation on real RCT-data
- If time: include image data

References

Chen, H., Aebersold, H., Puhan, M. A., and Serra-Burriel, M. (2025). Causal machine learning methods for estimating personalised treatment effects – insights on validity from two large trials.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):3–27.

Sick, B. and Dürr, O. (2025). Interpretable neural causal models with tram-dags. Accepted at the CLeaR 2025 Conference.

Sick, B., Hathorn, T., and Dürr, O. (2021). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2476–2481.