Proceedings of Machine Learning Research 275:1–25, 2025 4th Conference on Causal Learning and Reasoning

Interpretable Neural Causal Models with TRAM-DAGs

Beate Sick BEATE.SICK@UZH.CH
UZH, Zurich, Switzerland
ZHAW, Winterthur, Switzerland

Oliver Dürr OLIVER.DUERR@HTWG-KONSTANZ.DE
HTWG Konstanz, Germany
TIDIT.ch, Switzerland

Editors: Biwei Huang and Mathias Drton

Abstract

The ultimate goal of most scientific studies is to understand the underlying causal mechanism between the involved variables. Structural causal models (SCMs) are widely used to represent such causal mechanisms. Given an SCM, causal queries on all three levels of Pearl's causal hierarchy can be answered: $\mathcal{L}_1$ observational, $\mathcal{L}_2$ interventional, and $\mathcal{L}_3$ counterfactual. An essential aspect of modeling the SCM is to model the dependency of each variable on its causal parents. Traditionally this is done by parametric statistical models, such as linear or logistic regression models. This allows to handle all kinds of data types and fit interpretable models but bears the risk of introducing a bias due to the assumed rigid functional form. More recently neural causal models came up using neural networks (NNs) to model the causal relationships, allowing the estimation of nearly any underlying functional form without bias. However, current neural causal models are generally restricted to continuous variables and do not yield an interpretable form of the causal relationships. Transformation models range from simple statistical regressions to complex networks and can handle continuous, ordinal, and binary data. Here, we propose to use potentially deep TRAMs to model the functional relationships in SCMs allowing us to bridge the gap between interpretability and flexibility in causal modeling. We call this method TRAM-DAG and assume currently that the underlying directed acyclic graph (DAG) is known. For the fully observed case, we benchmark TRAM-DAGs against state-of-the-art statistical and NN-based causal models. We show that TRAM-DAGs are interpretable but also achieve equal or superior performance in queries ranging from $\mathcal{L}_1$ to $\mathcal{L}_3$ in the causal hierarchy. For the continuous case, TRAM-DAGs allow for counterfactual queries for three common causal structures, including unobserved confounding.

1. Introduction

Causal understanding is the ultimate goal in science and also essential in applications such as healthcare, economics, and policy-making because it allows to design effective interventions and make well-founded decisions. Structural Causal Models (SCMs) have become an established method for a mathematical representation of causal models. An SCM allows to tackle tasks on all three levels of Pearl's causal hierarchy: fitting observational distributions (Level 1), estimating interventional distributions (Level 2), and answering counterfactual queries (Level 3) (Pearl et al., 2000). One line of research in causal modeling is to estimate the directed acyclic graph (DAG), capturing the existence and directions of these mutual causal relationships as far as possible from observational data and asses if there are unobserved confounders. Another line of research starts from the DAG and focuses on estimating the functional form of the causal relationships of each variable on its

© 2025 B. Sick & O. Dürr.

causal parents. In this paper, we focus on this second line of causal modeling research and introduce TRAM-DAGs (see Fig. 1). The approaches for modeling the causal relationships can often be assigned to one of two choices: 1) A statistical approach that offers transparent parametric models for capturing causal relationships between variables of all data types but is prone to bias due to restrictive assumptions about functional forms (Peters et al. (2017)). 2) Neural network (NN) based causal models allowing for an unbiased estimation of complex relationships but suffering from their restriction to continuous data and their black box character.

To the best of our knowledge, our suggested TRAM-DAG is the first interpretable neural causal model that: a) can handle continuous, ordinal, binary or mixed data types and b) comprises classical statistical and NN-based approaches allowing to model the causal relationships of the SCM with interpretable or fully flexible model-parts within the same framework. We demonstrate that TRAM-DAGs achieve at least state-of-the-art performances in answering causal queries across the three levels of Pearl's causal hierarchy while retaining the interpretability required for understanding causal relationships.

2. Existing approaches for estimating causal relationships in SCMs

Estimating the functional relationships in SCMs can be broadly divided into statistical and NN-based approaches.

2.1. Causal models based on neural networks

There is a growing body of literature on estimating causal relationships using NNs, mostly generative neural network models, see Poinsot et al. (2024) for a recent comprehensive review. Theoretical results on the identifiability of neural causal models with continuous variables are discussed in Xia et al. (2023) for all three levels of causal hierarchy: fitting the observational data ($\mathcal{L}_1$), estimating intervention effects and interventional distributions ($\mathcal{L}_2$) and answering counterfactual questions ($\mathcal{L}_3$). These results are supplemented by numerical experiments using simple feed-forward NNs. Other methods go beyond simple feed-forward NNs. E.g. Sánchez-Martin et al. (2022) introduced Variational Graph Autoencoders (VACA). In VACA, the encoder graph network is not allowed to have hidden layers to allow for $\mathcal{L}_3$ identification, which limits the expressiveness of the approach. Handling $\mathcal{L}_2$ queries is also possible with sum-product network, see Poonia et al. (2024) or circuit models Wang and Kwiatkowska (2023). A particularly interesting class of causal models capable of $\mathcal{L}_3$ queries are models with a bijective generation mechanism (BGM) as described by Nasr-Esfahany et al. (2023). Their work demonstrated that models in this class are identifiable in the fully observed and two other cases with unobserved confounders, i.e. an instrumental variable or a backdoor setting. Specifically, a BGM model trained to fit continuous observational data at $\mathcal{L}_1$ can also predict $\mathcal{L}_2$ and $\mathcal{L}_3$ queries. However, to be bijective BGMs are restricted to continuous variables (see Section 4.2 for a discussion). Prominent members of that class of BGMs are normalizing flows (NFs). NFs rely on a single or a series of simple, invertible transformations to map variables to a simpler latent distribution - hence NF and TRAMs rely on the same idea (see Section 3.2). Initially, NFs have been proposed for causal estimation by Khemakhem et al. (2021), who introduced CAREFL that relies on chaining many simple transformations. Other recent NF-based methods achieve flexible transformations without chaining by directly modeling monotonic transformation. E.g., Balgi et al. (2024) uses unconstrained monotonic NNs. However, all current NF-based methods suffer from their black box character and are generally restricted to continuous data.

2.2. Causal models based on statistical models

Causal modeling dates back to the 1920s when path diagrams were introduced to describe causal relationships by Wright (1920). To describe the functional relationships in these causal models, structural equations like discussed in Section 3.1 were formally developed in the 1970s by Jöreskog (1970). Pearl introduced 1995 the do-calculus based on DAGs for answering causal queries Pearl (1995). While often linear regression models were used in the past, modern approaches can model more complex structures and use e.g. generalized linear models (GLMs) to set up the structural equations. Classical statistical models are usually not over-parametrized, and their interpretable parameters can be consistently estimated. A drawback of statistical modes are their limited flexibility, which can lead to suboptimal estimates of observational and interventional distributions. An overview of causal inference in statistics can be found in Pearl (2009) and a detailed discussion of complete identification methods for the causal hierarchy in Shpitser and Pearl (2008). In our study, we are in a quite easy setting because we assume a fully observed DAG, and utilize well-characterized transformation models to estimate the causal relationships. Hence, $\mathcal{L}_2$ queries can be solved by the do-calculus of Pearl when the observational data is accurately fitted Pearl (1995).

3. Background

We briefly introduce the necessary background needed for the proposed TRAM-DAG method (see Fig. 1) for causal modeling.

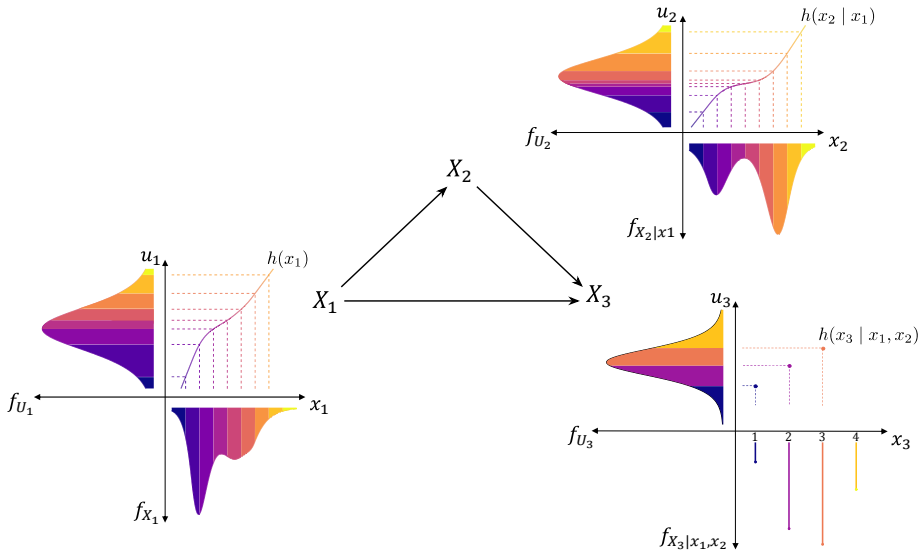

Figure 1: TRAM-DAG model: Each variable in the DAG is modeled by a TRAM. The TRAMs of the continuous variables X_1 and X_2 have a continuous h, and the TRAM of the discrete ordinal variable X_3 has a discrete h. For variables with parents, the conditional transformation function $h(X_i|\text{pa}(X_i))$ and outcome distribution $f_{X_i|\text{pa}(X_i)}$ depend on the values of the parents.

3.1. Background on DAGs and SCMs for TRAM-DAGs

In this study, we assume that the underlying causal structure, given by a directed acyclic graph (DAG) (see Fig. 2), is known. We focus on the case where all variables are observed. However, we would like to emphasize that for the continuous case, TRAM-DAGs are in the class of bijective generation models (BGM) (see Section 4.2) and therefore applicable beyond the full observed case Nasr-Esfahany et al. (2023). For the ease of discussion, we restrict to causal models where we have d mutual independent noise variables U_i meaning that we have no unobserved confounders. Although noise variables are typically omitted from DAGs for clarity, we include them in the DAG shown in Fig. 2.

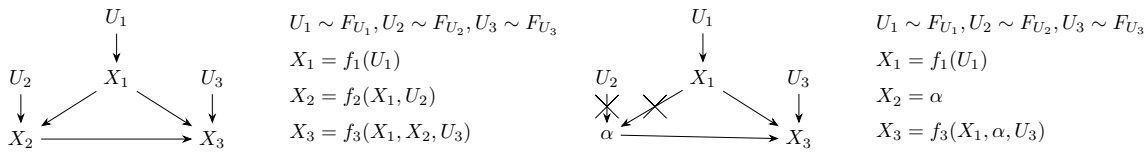

Figure 2: **Left:** DAG and corresponding SCM skeleton for three observed variables X_1, X_2, X_3 and unobserved noise U_1, U_2, U_3. **Right:** The post-interventional DAG and SCM skeleton when performing a do$(X_2 = \alpha)$ intervention.

Taken together, the collection of of d (structural) assignments,

$$X_i := f_i(\text{pa}(X_i), U_i), \quad \text{for } i = 1, \ldots, d. \tag{1}$$

and the specification of the d mutual independent noise distributions F_{U_i} define a structural causal model (SCM) for the involved variables $X_1, ..., X_d$. Please note that given the observed data, the form of the functions f_i in an SCM would change if another noise distribution F_{U_i} is assumed.

When performing a deterministic do-intervention in a causal model, for example, do$(X = \alpha)$, the intervened variable is forced to take the value of the do-intervention. Consequently, no other variables influence the intervened variable, resulting in a post-intervention DAG where all directed edges pointing to the intervened variable are removed (see right panel of Fig. 2). The post-interventional SCM is updated only for the intervened variable, which now takes on the fixed value imposed by the intervention.

3.2. Background on transformation models as needed for TRAM-DAGs

We use (deep) transformation models (TRAMs), which have so far only been used in non-causal regression tasks (Hothorn et al., 2014; Sick et al., 2021; Kook et al., 2022b; Baumann et al., 2021), for causal modeling by using the causal parents of a variable as predictors (Fig. 1). We can directly use the core idea of TRAMs (Hothorn et al., 2014) to construct the conditional distribution of a variable X_i in a causal model, whether continuous, ordinal, or binary, on its parents as follows: A strictly monotone increasing conditional transformation function $h(x_i \mid \text{pa}(x_i))$ is fitted that maps the unspecified conditional outcome distribution $F_{X_i|\text{pa}(X_i)}$ to a fixed continuous latent distribution F_u with a log-concave and continuous density f_u (see Fig. 1). This approach allows to model the

conditional outcome distribution $P(X_i \leq x_i|\text{pa}(X_i)) = F_{X_i|\text{pa}(X_i)}(x_i)$, as

$$F_{X_i|\text{pa}(X_i)}(x_i) = F_U(h(x_i|\text{pa}(x_i))). \tag{2}$$

Structure of the transformation function h: The transformation function $h = h_I + h_S$ for a target X_i consists of an intercept h_I that can depend on the parents $\text{pa}(X_i)$ and potentially a shift term h_S that can potentially consist of a sum of several linear and complex shift term that depend on the parents. Depending on the structure of h, the TRAM can be interpretable or be complex allowing for a high flexibility (see Section 3.2.1).

Choice of the latent distribution F_U: The interpretation scale of the shift terms in h depends on the choice of F_u, while F_u does not impact the ability to accurately estimate conditional outcome distributions (Hothorn et al., 2014). In this study, we always use the standard logistic distribution as F_u since it allows us to interpret the shift terms in h as log-odds-ratios (see Appendix A.1).

Intercept function for discrete ordinal or binary variables: For an ordinal variable X_i with levels $1, 2, ...K$, we use a monotone increasing discrete function to model the intercept function h_0 (see e.g. X_3 in Fig. 1). The discrete intercept function, which potentially depends on the parents, is given by:

$$h_I(X_i = k|\text{pa}(X_i)) = \vartheta_k(\text{pa}(X_i)) \tag{3}$$

where ϑ_k is a strictly monotone increasing sequence for $k = 1, ...K$. The probability for a class level $k \in \{2, ..., K-1\}$ is given by the area under the latent density over the interval $[\vartheta_{k-1}, \vartheta_k]$, for $k = 1$ the interval is $[-\infty, \vartheta_1]$, for $k = K$ it is $[\vartheta_{K-1}, \infty]$. A binary outcome can be seen as a special case where h only consists of one cut-point ϑ, cutting the area under latent density in two parts where the lower part represents $P(X_i = 0) = F_U(h(x_i = 0)$.

Intercept function for continuous variables: For continuous variables, we use Bernstein polynomials to model the continuous intercept function, which potentially depends on the parents.

$$h_I(x_i|\text{pa}(X_i)) = \frac{1}{M+1} \sum_{k=0}^{M} \vartheta_k(\text{pa}(X_i))\text{Be}_{k,M}(x_i), \tag{4}$$

where $\vartheta_k, k = 0, ..., M$ are strictly monotone increasing coefficients of the Bernstein polynomial to ensure a strictly monotone increasing h and $\text{Be}_{k,M}(x_i)$ denotes the density of a Beta distribution with parameters $k+1$ and $M-k+1$. We choose Bernstein polynomials because they can easily be restricted to be strictly monotonically increasing and provide theoretical guarantees for approximating any conditional continuous distribution arbitrarily well as long as the order M is sufficiently large Hothorn et al. (2014). For such a bijective h we can directly formulate $X_i = f_i(\text{pa}(X_i), U_i) = h^{-1}(U_i|\text{pa}(X_i))$.

3.2.1. FLEXIBLE AND INTERPRETABLE DEEP TRAMS: CI, SI-CS, SI-LS

To model **fully flexible** function f_i for a variable X_i in an SCM we allow the parameters ϑ_k in the intercept of h (discrete or continuous) to change with the value of the parents of X_i (see Eq. (3), Eq. (4)). We call this a complex intercept (CI) model, which provides maximal flexibility and can approximate any conditional outcome distribution arbitrarily well, as shown in Hothorn et al. (2014), where CI models are referred to as response-varying effect models.

To model a **causally interpretable** effect for each parent $X_j \in \text{pa}(X_i)$ on X_i, we design the transformation function as $h(x_i|\text{pa}(x_i)) = h_I(x_i) + \sum_j s(x_j)$ with a simple intercept (SI) h_I that does not depend on the parents, and additive interpretable shift terms $s(x_j)$ depending on the values x_j taken by the parents $X_j \in pa(X_i)$. A shift term $s(x_j)$ is either a linear shift (LS) $\text{LS}_{x_j} = \beta_j x_j$ or a complex shift (CS) $\text{CS}_{x_j} = \gamma(x_j)$. In this study, all shift parameters in the TRAM-DAG can be interpreted as log-odds-ratios since we use in all experiments the standard logistic distribution as F_U. While being the least flexible, the linear shift terms in the transformation h are the most interpretable. The parameter β_j can be causally interpreted as the log-odds ratio. Hence, $\exp(\beta_j)$ represents the factor by which the odds, $\text{odds}(X_i \leq x) = \frac{P(X_i \leq x)}{1-P(X_i \leq x)}$, change when intervening on the parent $X_j \in \text{pa}(X_i)$ by increasing it by one unit (see Appendix A). Importantly, in causal models, we do not require that all other parents $X_{j'} \in \text{pa}(X_i)$ with $j' \neq j$ stay constant; they may also change upon the intervention on X_j. Appendix C.4 provides an illustrative experiment showing that the causal parameter β_j, that was estimated on observational data, can be used to correctly predict the interventional effect of a parent X_j on the target X_i, by comparing e^{β_j} to the change of the $\text{odds}(X_i \leq x)$ when the intervention is actually performed in the DGP by increasing the parent X_j by one unit. For the more flexible complex shift terms $\text{CS}_{x_j} = \gamma(x_j)$, the change in the odds when increasing X_j by one unit can be expressed as $\exp(\gamma(x_j + 1) - \gamma(x_j))$. While the causal effect of a CS cannot be summarized by a single coefficient anymore, it can be interpreted by plotting $\gamma(x_j)$ against x_j (see e.g. Fig. 7).

To decide between a TRAM with full flexibility or interpretable effects, we follow the top-down approach as described in Hothorn (2018) aiming for maximal interpretability without sacrificing too much predictive performance as measured by the likelihood on an independent test set.

4. TRAM-DAGs

Here, we describe briefly how to fit TRAM-DAGs (see Fig. 1) and how to use them for tackling causal queries on all three levels in Pearl's causal hierarchy. The structure of TRAM-DAGs can be described by meta-adjacency matrix MA (see, e.g., Fig. 3) where the element in the i-th row and jth column describes the effect type of X_i on X_j which can be either a complex intercept (CI), a complex shift (CS), a linear shift (LS) or no influence at all (0). If the j-th column holds no CI-entry, then the intercept is modeled as SI, if all entries are 0, then X_j is a source node with $h(x_j) = SI$.

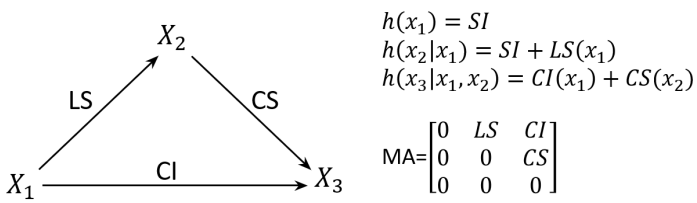

Figure 3: **Left**: DAG with meta information on the TRAMs, **Right:** h structures and the meta-adjacency matrix MA resulting from the DAG and the TRAM structures.

4.1. Architecture and training of TRAM-DAGs

Each (deep) TRAM in the TRAM-DAG can be trained separately on observational data. Details on training a (deep and interpretable) TRAM based on NNs are described in Kook et al. (2022b); Herzog et al. (2023); Kook et al. (2022a). For convenience, we have constructed a model that consists of a set of customized NNs including masked autoregressive flows (MAFs, see Papamakarios et al. (2018)) taking as input the meta-adjacency matrix MA (see Fig. 3) and the observational data. The output is the components of the d transformation functions, i.e. the ϑ-values for the intercept terms (see Eq. (3), Eq. (4)) and the linear and complex shift terms. We train all NNs jointly using the Adam optimizer.

4.2. Causal queries using TRAM-DAGs

If all variables are continuous and the TRAMs are flexible enough to accurately fit the observational data, we can show that our fitted TRAM-DAG can solve tasks on all three levels of Pearl's causal hierarchy.

Proposition: Counterfactual Equivalence of continuous TRAM-DAGs
Consider a SCM that contains only continuous variables. If the continuous and fully observed TRAM-DAG model can reproduce the observational distribution $\mathcal{L}_1$, then it will also reproduce the same interventional $\mathcal{L}_2$ and counterfactual $\mathcal{L}_3$ queries as the SCM.
Proof A continuous TRAM-DAG is based on transformation functions modeled by Bernstein polynomials with strictly monotone increasing coefficients. This ensures that each transformation is strictly monotone increasing. Therefore, continuous TRAM-DAGs fall into the class of BGMs of and the Lemma B.2 in Nasr-Esfahany et al. (2023) holds, stating the equivalence. ■

Although we focus on the fully observed case in this paper, it is important to note that continuous TRAM-DAGs, are BGMs and so theoretically capable of handling also some additional scenarios with unobserved confounders, as demonstrated in Lemma B.3 and B.4 in Nasr-Esfahany et al. (2023). Please also note that the monotonicity constraint does not effect the expressiveness of the transformation. In the following, we will show how tasks on $\mathcal{L}_1$ and $\mathcal{L}_2$ are tackled by all kinds of TRAM-DAGs and tasks of $\mathcal{L}_3$ by continuous TRAM-DAGs.

$\mathcal{L}_1$: Sampling from the observational distribution of continuous or mixed TRAM-DAGs To estimate the joint observational distribution, we sample d-dimensional observations $(x_1, ..., x_d)$ from our fitted TRAM-DAG. We first sample values u_j from each exogenous distribution $u_j \sim F_{U_j}, j = 1, ...d$. We go along the causal order and start with source nodes. In the case of a discrete variable X_i, we increase the sampled u_i to the next value of the discrete transformation function $h \geq u_i$. We deduce the corresponding sample x_i, by the requirement $u_i = h(x_i)$ which in case of continuous variables can be written as $x_i = h^{-1}(u_i)$. For non-source nodes X_j, $u_j = h(x_j|\text{pa}(x_j))$ is determined by the sampled values x_j of the parents. To get a sample x_j corresponding to u_j, we proceed as before.

$\mathcal{L}_2$: Estimating interventional distributions and treatment effects for continuous or mixed TRAM-DAGs We look at do-interventions where one variable is set to a certain value $\text{do}(X_i = \alpha)$. This results in a post-interventional graph where all arrows pointing to the intervened variables are deleted (see Figure 2). To estimate the joint interventional distribution, we proceed similarly as described in $\mathcal{L}_1$ but set $X_i = \alpha$ and go now along the causal order in the post-interventional DAG.

$\mathcal{L}_3$: Answering counterfactual queries with continuous TRAM-DAGs In a counterfactual task, we answer "what if" questions such as: What value would X_J have taken if the variable X_i had taken the values α instead of the observed value x_i? For illustration, look at Fig. 1 and assume $X_j = X_3$ and $X_i = X_2$ and that all variables are continuous. We answer the counterfactual question in three steps:
1) Abduction: determine the noise values u_i that correspond to the observations by $u_i = h(x_i|\text{pa}(X_i))$,
2) Action: Determine the counterfactual DAG for $x_i^c = \alpha$ that correspond to the post-interventional DAG (see for example Section 3.1),
3) Prediction: use the counterfactual DAG, the counterfactual value $x_i^c = \alpha$, the sampled noise values and go along the causal order to determine for all descendants X_j the counterfactual value as $x_j = h^{-1}(u_i|\text{pa}(x_j)$ by using the updated h where at least one parent is a descendant of X_i and has, therefore, an updated counterfactual value which likely resulted in a changed h. Note that for discrete or mixed TRAM-DAGs, counterfactual queries are not possible which is a fundamental limitation of causal models based on discrete ordinal variables, generated by interval censoring of an underlying continuous latent variable, and not a limitation of our proposed framework (see Appendix B for details).

5. Benchmarking TRAM-DAG against Neural Causal Models

Here, we benchmark TRAM-DAG with state-of-the-art NN and NF-based causal models. These models focus on estimating observational, interventional, and counterfactual distributions without requiring the interpretability of the fitted SCM. The code to reproduce the experiments can be found at: https://github.com/tensorchiefs/tram-dag.

5.1. Observational Distribution $\mathcal{L}_1$

To illustrate TRAM-DAG's ability to fit complex observational continuous data flexibly, we replicate an example originally introduced in Sánchez-Martin et al. (2022) and fitted with their VACA method. The data generating process (DGP) consists of three variables, X_1, X_2, X_3, details are provided in Appendix C.1. The variable X_1 follows a bimodal distribution, and X_2 and X_3 are linearly dependent on X_1. This leads to non-Gaussian marginal distributions also for X_2 and X_3 (see diagonal in Fig. 4). We benchmarked flexible TRAM-DAG with Causal Normalizing Flow (CNF) Javaloy et al. (2024). Fig. 4 shows that the distribution of the samples from the fitted TRAM-DAG model closely resamples the distribution of the DGP, including the bimodal distribution of X_1, while the sample distribution from the fitted CNF fails to capture the bimodal distribution of X_1, likely due to the inflexibility of the used transformation. We then adapted Neural Spline Flows (NSF) (Durkan et al., 2019) to the causal setting and observed a similar performance (see Fig. 12) as achieved by TRAM-DAGs. This highlights the importance of achieving flexibility in distribution modeling, which can be easily achieved with TRAM-DAGs.

5.2. Interventional Distribution $\mathcal{L}_2$

Next, we benchmark TRAM-DAG's capability to model interventional distributions by replicating the interventional experiment in Javaloy et al. (2024). We used the fitted complex intercept TRAM-DAG to perform do-interventions on X_2 with $\text{do}(X_2) = -3$, $\text{do}(X_2) = -2$, and $\text{do}(X_2) = 0$. Comparing the ground truth interventional distribution with the estimated interventional distribution

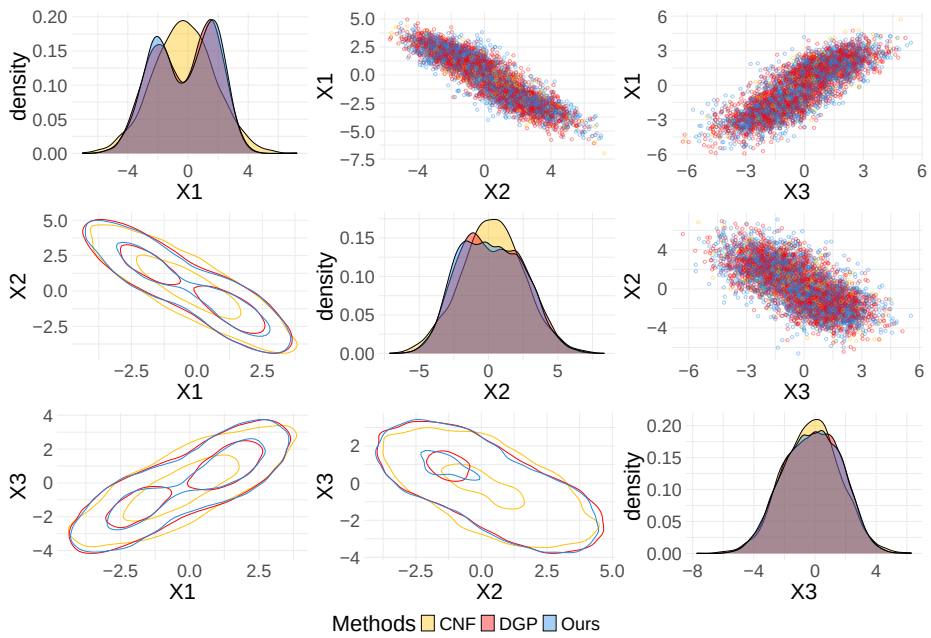

Figure 4: $\mathcal{L}_1$: Comparative analysis of joint and marginal observational distributions for samples generated by the DGP (see Appendix C.1 red), our fitted TRAM-DAG (blue), and the fitted CNF from the original study (yellow) (Javaloy et al. (2024)). The diagonal shows estimates of the marginal distributions, the lower triangle shows 2D density estimates, and the upper triangle presents scatter plots with subsampling.

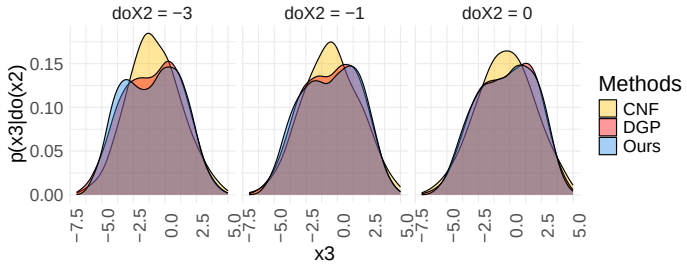

Figure 5: $\mathcal{L}_2$: Interventional Distribution $P(X_3|\text{do}(X_2))$ for different values of the do-intervention resulting from the DGP (see Appendix C.1 red), our fitted TRAM-DAG (blue), and the fitted CNF from the original study (yellow) (Javaloy et al. (2024)).

achieved with the CNF method and our TRAM-DAG method Fig. 5 demonstrates TRAM-DAG's ability to capture the correct and complex interventional distribution and outperforms CNF. Please note that it is straightforward to estimate from the interventional distribution the treatment effect of increasing X_2 by one unit from -3 to -2 by the difference of the means of the estimated distributions, $\mathbb{E}(X_3|\text{do}(x_2 = -2)) - \mathbb{E}(X_3|\text{do}(x_2 = -3)$.

5.3. Counterfactual queries $\mathcal{L}_3$

To evaluate TRAM-DAG's performance on counterfactual queries, we replicate the counterfactual experiments presented in CAREFL (Khemakhem et al., 2021) based on a non-linear DGP with four variables (see Appendix C.2 for more details). Following Khemakhem et al. (2021), we focus on the DGP generated observations, from which we pick the following observation $x_{\text{obs}} = (2.00, 1.50, 0.81, -0.28)$. We then consider two counterfactual queries:

(i) What would the expected value of x_3 would have been, if the variable X_2 would have taken the values $x_2 = \alpha$ instead of observed value $x_2 = 1.5$?

(ii) What would the expected value of x_4 would have been, if X_1 would have taken the value $x_1 = \alpha$ instead of the observed $x_1 = 2$?

In both experiments α values in the range between -3 and 3 are considered (see Fig. 6). The

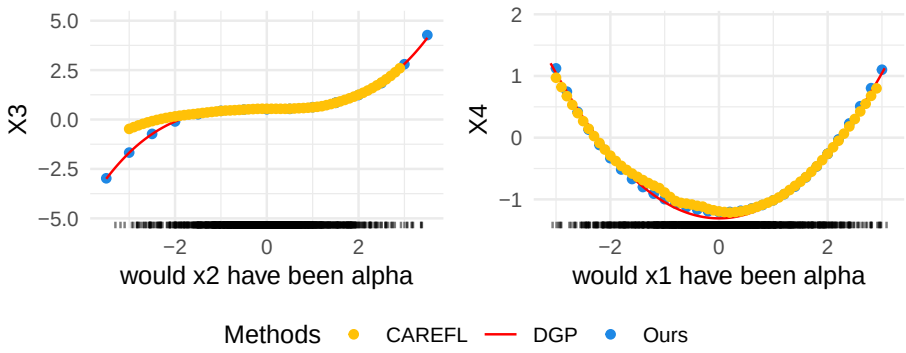

Figure 6: $\mathcal{L}_3$: Results of the counterfactual queries as posed in CAREFL based on a picked four-dimensional observation (Khemakhem et al., 2021). **Left:** the expected value of the counterfactual distribution of X_3 (depicted on the y-axis) if X_2 would have taken the value α (depicted on the x-axis) instead of the observed value $x_2 = 1.5$. **Right:** the expected value of the counterfactual distribution of X_4 if X_1 would have been α instead of the observed value $x_1 = 2$. Shown are results from the DGP (see Appendix C.2 red), our TRAM-DAG (blue) and CAREFL (yellow).

results of this benchmark experiment show that our TRAM-DAG closely resamples the true counterfactual results and slightly outperforms the CAREFL method.

Overall, TRAM-DAG is on par or slightly outperforms state-of-the-art NN- and NF-based causal methods on all three levels of Pearl's hierarchy.

6. Experiments with interpretable components

The following experiments focus on demonstrating that an SCM where the causal relationships are given by interpretable functions can be fitted by TRAM-DAGs without losing the interpretability. That is because causal TRAMs can be set up as interpretable models allowing the user to understand and judge the modeled causal effect of each parent on the target variable, as well as predicting the effect of interventions (see Section 3.2.1, Appendix C.4).

6.1. Continuous Case

We start with an SCM involving three continuous variables (see DAG in Fig. 7).

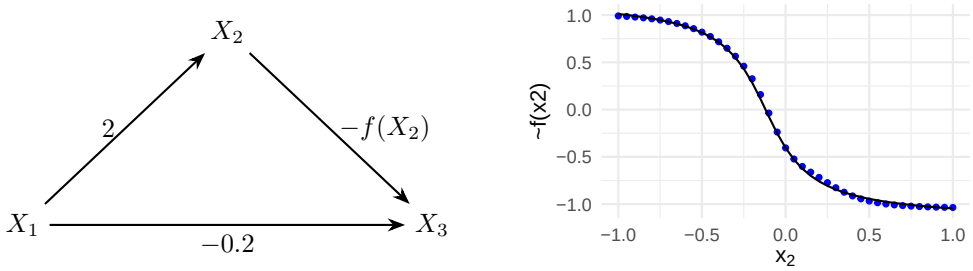

Figure 7: **Left:** DAG with three continuous variables and meta information for shift terms in h: $h(x_1) = SI$, $h(x_2|x_1) = SI + 2 \cdot x_1$, $h(x_3|x_1, x_2) = SI - 0.2 \cdot x_1 - f(x_2)$
Right: Complex shift term $CS = -f(x_2)$ depicted as solid black line for the DGP ground truth $f(x_2) = 0.75 \cdot \arctan(5 \cdot (x_2 + 0.12))$ along with blue dots for the estimated CS in the fitted TRAM-DAG.

The DGP generates all three variables using a TRAM model with simple intercept and interpretable shift terms. We now perform three experiments, each with slightly different DGP and TRAM-DAG specifications. In the DGPs of all three experiments we use linear effects of X_1 on X_2 and X_3 ($\beta_{12} = 2$, $\beta_{13} = -0.2$) and also use linear shift term to specify these effects in the fitted TRAM-DAGs. However, we use different choices for the functional form f of the causal effect of X_2 on X_3 in the DGP (see e.g. Fig. 7) and different specifications in the fitted TRAM-DAGs.

Linear-shift DGP and linear-shift model We use in the DGP $f(X_2) = -0.3 \cdot X_2$, resulting in in $h(x_3|x_1, x_2) = SI + \beta_{13} \cdot x_1 - f(x_2) = SI - 0.2 \cdot x_1 + 0.3 \cdot x_2$ and fit a correctly specified TRAM-DAG model with linear shift terms for all variables. The estimated coefficients are in good agreement with the true values from the DGP ($\beta_{12} = 2$, $\hat{\beta}_{12} = 1.98$; $\beta_{13} = -0.2$, $\hat{\beta}_{13} = -0.21$; $\beta_{23} = 0.3$, $\hat{\beta}_{23} = 0.26$) and can be causally interpreted as log-odds-ratio (see Section 3.2.1). The fitting process in shown in Appendix C.3.

Complex-shift DGP and complex-shift model Next, we increase the complexity of the DGP by defining $f(X_2) = 0.75 \cdot \arctan(5 \cdot (X_2 + 0.12))$ which introduces a non-linear causal impact of X_2 on X_3. We fit correctly specified TRAM-DAG with a complex shift term and achieve correctly estimated coefficients in the linear shift terms of X_1 on X_2 and X_3 ($\beta_{12} = 2$, $\hat{\beta}_{12} = 2.07$; $\beta_{13} = -0.2$, $\hat{\beta}_{13} = -0.203$) and a well fitted CS function $f(X_2)$ (see right panel in Fig. 7) confirming TRAM-DAG's capability to capture complex non-linear dependencies.

Linear-shift DGP and complex-shift model Now we use in the DGP again a linear effect of X_2 on X_3 with $f(X_2) = -0.3 \cdot X_2$. However, we fit a miss-specified TRAM-DAG model with a complex shift term to model the effect of X_2 on X_3. Even under such a misspecification of the TRAM-DAG, the linear form of $f(x2)$ is approximately matched, and the coefficients $\beta_{12} = 2$, $\beta_{13} = -0.2$ are well estimated (see Appendix C.3.3). Noteworthy, the observational and interventional distributions are accurately estimated (see Fig. 17).

Additional results of these experiments can be found in Appendix C.3.

6.2. Mixed data types

We now use an SCM involving three variables of mixed data types (see Fig. 8). Compared to the DGP from the continuous case in the last subsection we have replaced the continuous X_3 with an ordered categorical variable $X_3 \in \{1, 2, 3, 4\}$ (as in Fig. 1) and now use a positive linear effect of X_1 on X_3 ($\beta_{13} = 2$) and a negative effect ($\beta_{23} = -0.3$) if f is linear.

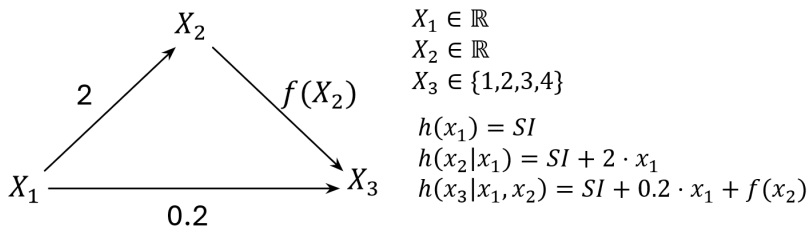

Figure 8: DAG and setting for the interpretation experiment in the mixed case.

.

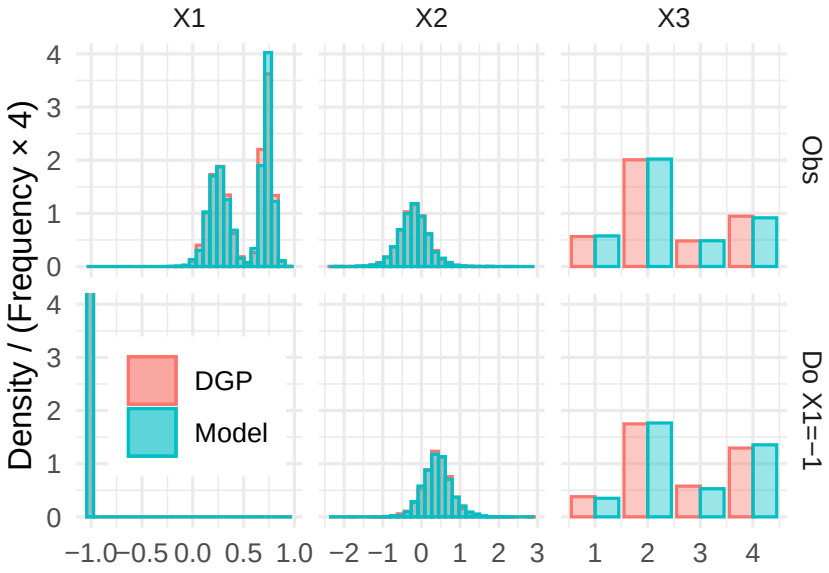

Figure 9: $\mathcal{L}_1$ and $\mathcal{L}_2$ for the mixed data experiment: Comparison of observational and interventional distributions where all shift therms were linear in the DGP and the fitted TRAM-DAG. The frequencies of X_3 have been multiplied by a factor of 4 for visual convenience.

In a first experiment with mixed data types we used for $f(x_2) = -0.3 \cdot x_2$. This results in $h(x_3|x_1, x_2) = SI + \beta_{13} \cdot x_1 + f(x_2) = SI - 0.2 \cdot x_1 - 0.3 \cdot x_2$ and showed that the coefficients in all three linear shift terms are accurately fitted (see Fig. 19). Then we showed that that observational and interventional distributions are accurately estimated in case of LS and CS (see (Fig. 9, Fig. 20).

We have added an illustrative example in Appendix C.4 to demonstrated that the estimated coefficients in the linear shift terms of a correctly specified and fitted TRAM-DAG allow to correctly predict the effect of an intervention on a parent variable in terms of the resulting change of the odds to observe values below or equal to a freely specifiable cutoff after the intervention compared to before.

With these experiments we have demonstrate TRAM-DAG's applicability to tackle $\mathcal{L}_1$ and $\mathcal{L}_2$ tasks with mixed data types.

7. Conclusion

In this paper, we introduced TRAM-DAGs, a novel framework for interpretable neural causal models. TRAM-DAGs range from transparent and interpretable causal models to causal models with the flexibility of deep learning models. Tuning the level of interpretability and flexibility for certain applications depends on the complexity of the data and the needed interpretability. Continuous TRAM-DAGs can be trained using observational data and used to answer queries across all three levels of Pearl's causal hierarchy: observational($\mathcal{L}_1$), interventional ($\mathcal{L}_2$), and counterfactual ($\mathcal{L}_3$). Mixed TRAM-DAGs are restricted to queries within the first two levels ($\mathcal{L}_1$ and $\mathcal{L}_2$). The possibility to incorporate binary, ordinal, continuous, or mixed data types in a TRAM-DAG is a big advantage compared to other state of the art causal models that rely on NNs. In the continuous case, TRAM-DAGs fall within the class of bijective generation mechanism (BGM) models, inheriting all BGM properties, particularly their applicability to common causal structures with unobserved confounding (Nasr-Esfahany et al., 2023).

ACKNOWLEDGEMENTS

We want to thank Lucas Kook for helpful discussions and Pascal Bühler for his help with the figures. We sincerely thank the reviewers for their valuable feedback. This work was partially supported by Carl-Zeiss-Stiftung in the project "DeepCarbPlanner" (grant no. P2021-08-007).

References

Sourabh Balgi, Adel Daoud, Jose M Pena, Geoffrey T Wodtke, and Jesse Zhou. Deep learning with dags. *arXiv preprint arXiv:2401.06864*, 2024.

Philipp FM Baumann, Torsten Hothorn, and David Rügamer. Deep conditional transformation models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 3–18. Springer, 2021.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural Spline Flows. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lisa Herzog, Lucas Kook, Andrea Götschi, Katrin Petermann, Martin Hänsel, Janne Hamann, Oliver Dürr, Susanne Wegener, and Beate Sick. Deep transformation models for functional outcome prediction after acute ischemic stroke. *Biometrical Journal*, 65(6):2100379, 2023.

Torsten Hothorn. Top-down transformation choice. *Statistical Modelling*, 18(3–4):274–298, 2018. doi: 10.1177/1471082X17748081.

Torsten Hothorn, Thomas Kneib, and Peter Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):3–27, 2014. doi: 10.1111/rssb.12017.

Adrian Javaloy, Pablo Sanchez-Martıin, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36, 2024.

Karl G Jöreskog. A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2):i–41, 1970.

Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal Autoregressive Flows. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, March 2021.

Lucas Kook, Philipp FM Baumann, Oliver Dürr, Beate Sick, and David Rügamer. Estimating conditional distributions with neural networks using r package deeptrafo. *arXiv preprint arXiv:2211.13665*, 2022a.

Lucas Kook, Lisa Herzog, Torsten Hothorn, Oliver Dürr, and Beate Sick. Deep and interpretable regression models for ordinal outcomes. *Pattern Recognition*, 122:108263, 2022b. doi: 10.1016/j.patcog.2021.108263.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *International conference on machine learning*, pages 25733–25754. PMLR, 2023.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, 2018. URL https://arxiv.org/abs/1705.07057.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Audrey Poinsot, Alessandro Leite, Nicolas Chesneau, Michèle Sébag, and Marc Schoenauer. Learning Structural Causal Models through Deep Generative Models: Methods, Guarantees, and Challenges, May 2024. URL http://arxiv.org/abs/2405.05025. arXiv:2405.05025 [cs, stat].

Harsh Poonia, Moritz Willig, Zhongjie Yu, Matej Zečević, Kristian Kersting, and Devendra Singh Dhami. χ spn: Characteristic interventional sum-product networks for causal inference in hybrid domains. *arXiv preprint arXiv:2408.07545*, 2024.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. 2008.

Beate Sick, Torsten Hathorn, and Oliver Dürr. 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. IEEE, 2021.

Pablo Sánchez-Martin, Miriam Rateike, and Isabel Valera. VACA: Designing Variational Graph Autoencoders for Causal Queries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):8159–8168, June 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i7.20789. URL https://ojs.aaai.org/index.php/AAAI/article/view/20789. Number: 7.

Benjie Wang and Marta Kwiatkowska. Compositional probabilistic and causal inference using tractable circuit models. In *International Conference on Artificial Intelligence and Statistics*, pages 9488–9498. PMLR, 2023.

Sewall Wright. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, 6(6):320–332, 1920.

Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The eleventh international conference on learning representations*, 2023.

SUPPLEMENTARY MATERIAL

Appendix A. Transformation models

TRAMs were introduced in 2014 as a flexible distributional regression method for tabular ordered data which can be continuous, discrete, or censored. (Hothorn et al., 2014). Later, TRAMs were extended to deep TRAMs by Sick et al. (2021) using neural networks, allowing the inclusion of unstructured data modalities like images. TRAMs comprise most classical statistical regression models, like linear or logistic regression or other GLMs, and have hence the same interpretability of their parameters and the same guarantees as these well-established statistical models (Hothorn et al., 2014). However, TRAMs do provide a much larger family of models since TRAMs do not require pre-specify the family of the outcome distribution and allow to model flexible outcome distributions that change with the predictors, resulting in distributions that do not even need to belong to a known distribution family.

A.1. Interpretability of the shift terms

The choice of the latent distribution F_u has no influence of the prediction power of the TRAM but determines the interpretation scale of the shift terms in the transformation function h (Hothorn et al., 2014). In our experiments we always use the standard logistic distribution $P(Y \leq y) = F_Y(y) = F_{SL}(z) := (1+\exp(-z))^{-1}$ as latent distribution with inverse $F_{SL}^{-1}(P) = \log\left(\frac{P}{1-P}\right) = \log(\text{odds})$ that allows to interpret the shift parameters as log-odds ratios. This is known from the logistic regression where the target Y is binary, but is also valid for ordinal or continuous target variables as demonstrated here for a $SI - LS_{x1} - CS_{x2}$ model with $h(y|x_1, x_2) = h_0(y) + \beta_1 x_1 + \gamma(x_2)$ and a continuous target Y:

$$
\begin{aligned}
F_{Y|x_1,x_2}(y) &= F_{SL}(h(y|x_1, x_2)) \\
\Leftrightarrow P(Y \leq y|x_1, x_2) &= F_{\text{SL}}(h_0(y) + \beta_1 x_1 + \gamma(x_2)) \\
\Leftrightarrow \log(\text{odds}(Y \leq y|x_1, x_2) &= h_0(y) + \beta_1 x_1 + \gamma(x_2) \\
\text{OR}_{x_1 \to x_1+1} &= \frac{\text{odds}(Y \leq y|x_1+1, x_2)}{\text{odds}(Y \leq y|x_1, x_2)} = \frac{\exp\left(h_0(y) + \beta_1(x_1+1) + \gamma(x_2)\right)}{\exp\left(h_0(y) + \beta_1 x_1 + \gamma(x_2)\right)} \\
&= \exp(\beta_1) \\
\text{OR}_{x_2 \to x_2+1} &= \frac{\text{odds}(Y \leq y|x_1, x_2+1)}{\text{odds}(Y \leq y|x_1, x_2)} = \frac{\exp\left(h_0(y) + \beta_1 x_1 + \gamma(x_2+1)\right)}{\exp\left(h_0(y) + \beta_1 x_1 + \gamma(x_2)\right)} \\
&= \exp(\gamma(x_2+1) - \gamma(x_2))
\end{aligned}
$$

Hence $\exp(\beta_i)$ is interpreted as odds-ratio, which is the factor by which the odds for $Y \leq y$ is changing if increasing x_i by one unit and holding all other variables constant. Remarkably, this holds for any threshold value y, and hence, these models are called proportional odds models. Equivalently, the parameter β_i in a LS term can be interpreted as log-odds-ratio $\beta_i = \log(\text{OR}x_i \to x_i + 1)$. Please note that in a causal model, where all predictors X_i are direct causal parents of the target Y, this interpretation holds causally. This means that when intervening on the predictor X_i by increasing it by one unit, the other parents of Y may change upon this intervention and the observed odds of $Y \leq y$ in the interventional data will differ by the factor $\exp(\beta_i)$ compared to the observational data. We demonstrate that β in a LS does correctly predict this interventional

effect in an illustrative example in Appendix C.4. This requires that the causal model is correct, meaning it matches the data-generating process as in most of our experiments.

Proportional odds models are more commonly used for an ordinal target where the parameters in the LS terms quantify the change of the odds for $Y \leq y_k$ holding for all class levels y_k.

The same math works for a binary target where we look at the odds for $Y \leq 0$, which is same as the odds for $Y = 0$. The odds for $Y = 0$ changes by the factor $\exp(\beta_i)$ if x_i is increased by one unit and all other predictors stay constant. Note that in most implementations of the logistic regression the default latent distribution is the standard logistic distribution and a β_i in the linear regression can be interpreted as log-odds-rations for $Y = 1$.

For a CS term, the difference of the estimated shifts can be interpreted as log-odds-ratio $\gamma(x_j + 1) - \gamma(x_j) = \log(\text{OR} x_j \to x_j + 1)$

Appendix B. Impossibility of Counter-Factual queries for discrete targets

Here we show why counterfactual queries can in general not be answered for discrete variables that are generated by censoring an underlying continuous variable. For example sport grades in an 100 meter running test are ordinal and give an incomplete quantification of the student's speed since all students who run the 100 meter in a certain interval of time get the same grade - the grades are interval censored.

Counterfactual queries typically require a unique mapping from the observed outcome $\mathbf{x}$ to the underlying noise realizations $\mathbf{u}$ ("abduction"), so that one can subsequently "re-run" $\mathbf{u}$ under a hypothetical intervention ("action" and "prediction"). For continuous outcome variables, this mapping can be made bijective under mild assumptions, rendering counterfactual queries well-defined. However, for discrete (e.g., binary or ordinal) outcome variables, the mapping from a continuous noise variable to the observed discrete value is necessarily many-to-one: an entire interval of latent noise values collapses to a single discrete outcome.

For illustration, look at Figures 1 and 10. Let's take the following counterfactual query: What value would X_3 have taken if the variable X_2 would have taken the values α instead of the observed value x_2? For the discrete variable X_3, the discrete transformation $h(x_3|\text{pa}(x_3))$ strictly monotone but not bijective. Hence, it is not possible to determine unambiguous values for the noise variable U_3 as needed in the abduction step of a counterfactual analysis . Imagine the observed level of X_3 was level two, $x_3 = 2$ (purple), then all noise values $u_3 \in [h(1|x_1, x_2), h(2|x_1, x_2)]$ (indicated by the purple bar) would be possible since all these noise values lead to the observed value $x_3 = 2$. This can then lead to problems in the prediction step. In a counterfactual situation we imagine that X_2 would have taken $x_2^c = \alpha$ instead of the observed x_2. Hence, the transformation function $h(x_3|x_1, \alpha)$ (indicated by crosses in the Figure) has probably changed compared to the original $h(x_3|x_1, x_2)$ (indicated by dots in the Figure). Then it can happen that $h(1|x_1, \alpha) \in [h(1|x_1, x_2), h(2|x_1, x_2)]$ (as illustrateted in the Figure), resulting in $x_3^c = 1$, if $u_3 \leq h(1|x_1, \alpha)$ and $x_3^c = 2$, if $h(1|x_1, \alpha) \leq u_3 \leq h(2|x_1, \alpha)$. As demonstrated in this example, it is not in general possible to determine an unambiguous counterfactual x_3^c for ordinal variables.

Appendix C. Additional Experimental Results and details of the experiments

The code to reproduce the experiments is available at: https://github.com/tensorchiefs/tram-dag

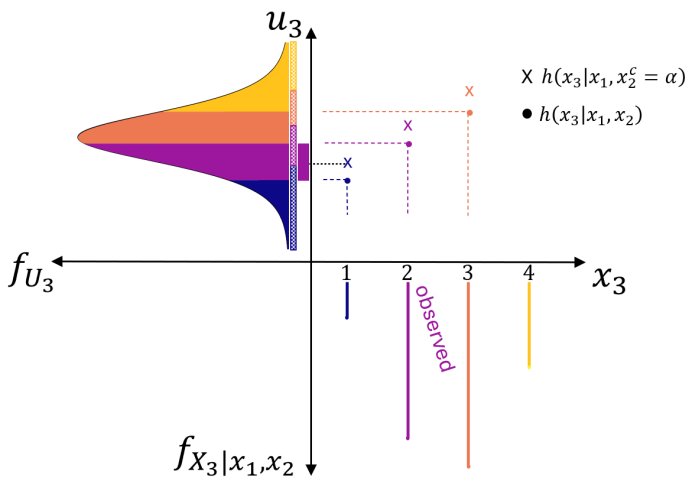

Figure 10: TRAM with discrete target X_3 with a original h corresponding to the observed values x_1, x_2 (indicated with dots) and a counterfactual h in the counterfactual situation $x_2^c = \alpha$ (indicated as crosses). The solid colored areas in f_{U_3} correspond to the cutpoints of the original h, while the counterfactual h would lead to shifted cutpoints as indicated by the hatched colored bars below f_{U_3}. If $x_3 = 2$ was observed then the corresponding noise value could have been any value in the following interval $u_3 \in [h(1|x_1, x_2), h(2|x_1, x_2)]$ indicated by the thick purple bar. In the counterfactual situation where X_2 would have taken the value α the unambiguity of u_3 results in an unambiguity of the counterfactual value x_3^c, because different parts of the possible noise values (solid purple bar) fall into different counterfactual bins (hashed bars) which would result in different counterfactual values of x_3^c.

C.1. Observational Distribution and do interventions

C.1.1. Data Generating Process

The Data Generating Process (DGP) follows the original VACA paper (see appendix E.1 in Sánchez-Martin et al. (2022)).

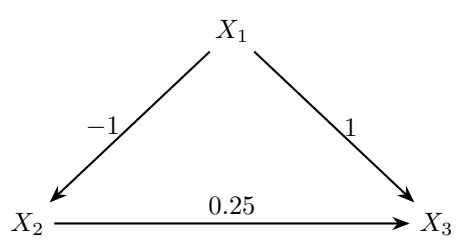

Figure 11: DAG of the DGP process in the original VACA paper

$$X_1 = \begin{cases} \mathcal{N}(-2, \sqrt{1.5}) \text{with probability } 0.5, \\ \mathcal{N}(1.5, 1) \text{with probability } 0.5 \end{cases}$$

$$X_2 = -X_1 + \mathcal{N}(0,1)$$
$$X_3 = X_1 + 0.25 \cdot X_2 + \mathcal{N}(0,1)$$

Note that X_1 follows a bimodal distribution according to the following DGP

C.1.2. MODELS

We compare our TRAM-DAG against default implementation of Causal Normalizing Flow (CNF) by Javaloy et al. (2024). The CNF is based on MAF-like NN with 3 hidden layers each of dimension 16 using affine linear transformations but is still not able to fit a bimodal observational or interventional distribution (see Fig. 4, Fig. 5). It is important to note that Javaloy et al. (2024) uses a different version of the DGP compared to Sánchez-Martin et al. (2022), where the bimodal distribution for X_1 is replaced by a standard normal distribution $\mathcal{N}(0,1)$ which could be fitted by CNF model.

C.1.3. ADDITIONAL EXPERIMENTS WITH NEURAL SPLINE FLOWS

In Figure 12, we replicate the experiment from Figure 4, but replace the inflexible Causal Normalizing Flow (CNF) presented by Javaloy et al. (2024) with a Neural Spline Flow (NSF). As expected, the NSF achieves a more accurate fit to the bimodal distribution of X_1, underscoring the value of using flexible transformations, such as NSF or TRAM-DAGs with complex intercepts modeled by Bernstein polynomials for modeling complex distributions.

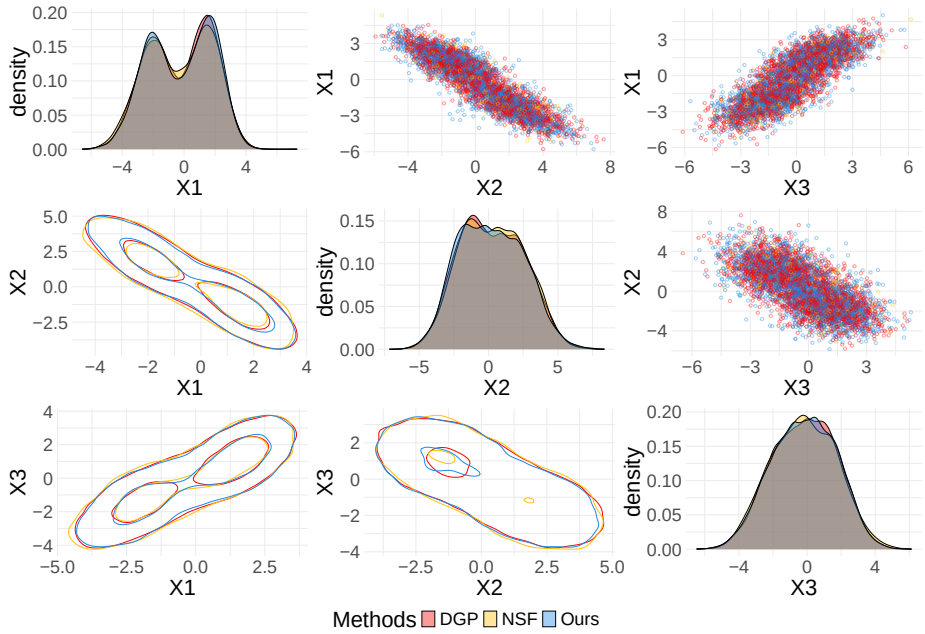

Figure 12: $\mathcal{L}_1$: Samples generated by the DGP (see Fig. 11), our fitted TRAM-DAG and fitted Neural Spline Flow (NSF). Same experiment as for Figure 4 but this time the inflexible CNF has been replaced by a flexible NSF.

C.2. Counterfactual CAREFL experiment

Here we present additional information for the counterfactual CARELF experiment done by Khemakhem et al. (2021). The DAG is depicted in the following causal graph in Fig. 13. The DGP for

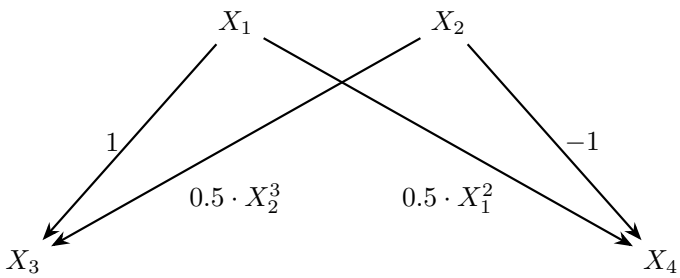

Figure 13: DAG of the DGP used in the counterfactual experiment to benchmark TRAM-DAG with the CAREFL method presented in 13

the four variables X_1, X_2, X_3, and X_4, where X_3 and X_4 holds as non-linear transformations of X_1 and X_2 and is defined as follows:

$$X_1, X_2 \sim \text{Laplace}(0, \frac{1}{\sqrt{2}})$$
$$X_3 = X_1 + 0.5 \cdot X_2^3 + \text{Laplace}(0, \frac{1}{\sqrt{2}})$$
$$X_4 = -X_2 + 0.5 \cdot X_1^2 + \text{Laplace}(0, \frac{1}{\sqrt{2}})$$

For our implementation, we employed the same architecture and training procedure as described in Section 4.1. We compare our results against the original CAREFL model, as presented in Figure 5 of Khemakhem et al. (2021).

C.3. Interpretable Experiments Continuous Case

Here, we give additional results for the experiments with DGP with three continuous variables and fitted interpretable TRAM-DAGs (see Section 6.1, Fig. 7).

For the simple intercept in the interpretable continuous TRAM-DAG, we used Bernstein polynomials of order $M = 20$ (see Eq. (4)). The training was conducted for 500 epochs on a dataset containing 40000 samples, utilizing the Adam optimizer with the default learning rate of 0.001.

C.3.1. LINEAR-SHIFT DGP AND LINEAR-SHIFT MODEL

The causal effect of X_2 on X_3 is in the DGP is linear given by $f(X_2) = -0.3 \cdot X_2$, resulting in in $h(x_3|x_1, x_2) = SI + \beta_{13} \cdot x_1 - f(x_2) = SI - 0.2 \cdot x_1 + 0.3 \cdot x_2$ and hence $\beta_{23} = 0.3$. The used TRAM-DAG models the causal impact of all parents on their target as linear shift term. Fig. 14 illustrates the evolution of the estimated coefficients throughout the training process, showing how the estimated coefficients in LS-terms of the TRAM-DAG converge towards the true coefficients of the DGP $\beta_{12} = 2$, $\beta_{13} = -0.2$, and $\beta_{23} = 0.3$. Please note, that the fitted TRAM-DAG accurately recovers the ground truth coefficients.

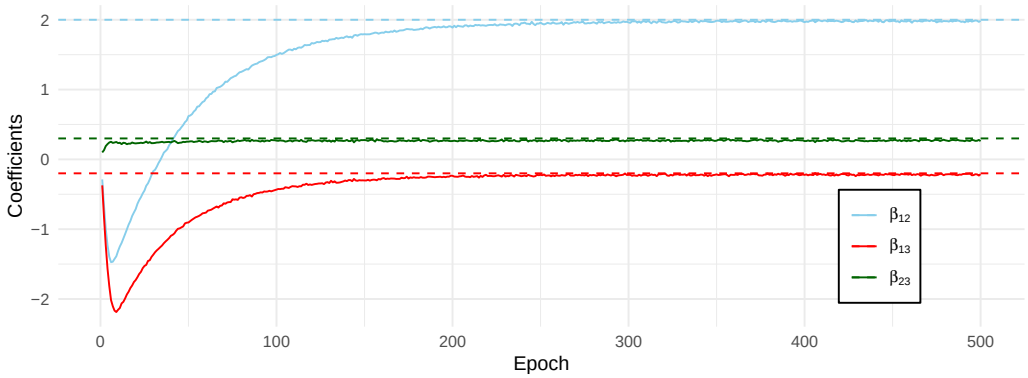

Figure 14: Interpretable continuous case experiment with three linear shift terms: Estimated coefficients β_{12}, β_{13}, and β_{23} over training epochs, with dashed lines indicating the true coefficient values of the DGP.

C.3.2. COMPLEX-SHIFT DGP AND COMPLEX-SHIFT MODEL

The causal effect of X_2 on X_3 in the DGP is modeled as complex shift $f(x_2) = 0.75 \cdot \arctan(5 \cdot (x_2 + 0.12))$. The other causal effects in the DGP remain linear effects.

Figure Fig. 15 shows that the estimated coefficients in LS-terms of the TRAM-DAG converge towards the true coefficients of the DGP $\beta_{12} = 2$, $\beta_{13} = -0.2$.

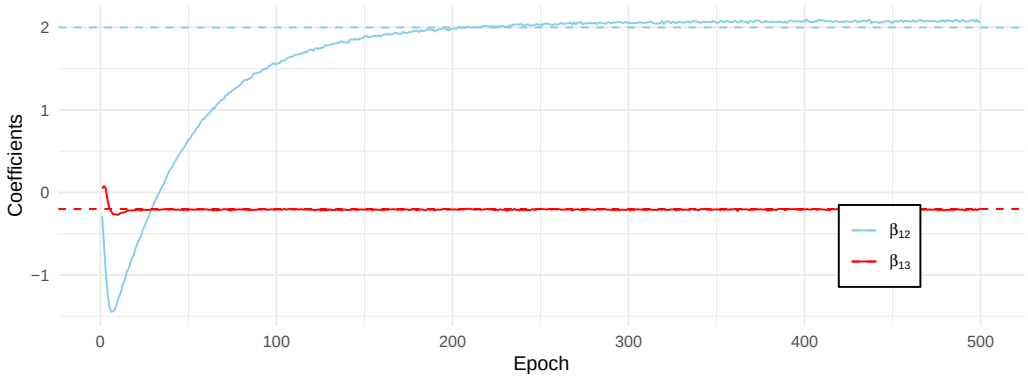

Figure 15: Interpretable continuous case experiment with two linear shift terms and one complex shift term: Estimated coefficients β_{12}, β_{13} of the two LS over training epochs, with dashed lines indicating the true coefficient values of the DGP.

In Fig. 16, we see that the estimated observational and interventional distributions match the corresponding distributions produced by the DGP.

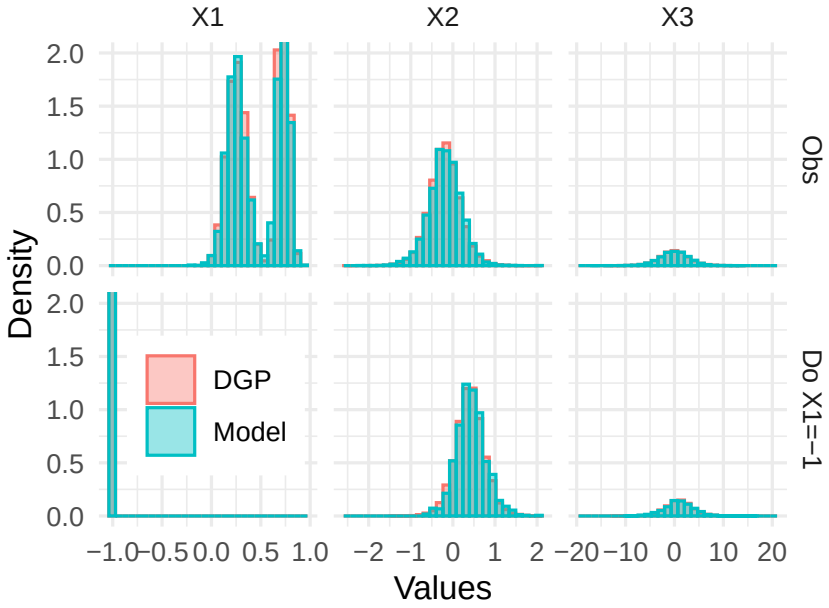

Figure 16: $\mathcal{L}_1$ and $\mathcal{L}_2$ in the interpretable continuous case experiment with two linear shift terms and one complex shift term: Comparison of observational distributions (upper panel) and interventional distributions (lower panel) in the continuous case as generated by the DGP or the fitted TRAM-DAG. The DGP holds $f(x_2) = 0.75 * \text{atan}(5 * (x_2 + 0.12))$ and the TRAM-DAG modeled this effect by a CS.

.

C.3.3. LINEAR SHIFT DGP, COMPLEX SHIFT MODEL

We now examine the scenario of a misspecified TRAM-DAG is fitted. The causal effect of X_2 on X_3 is in the DGP is linear $\beta_{23} = 0.3$, the other causal effects in the DGP are also linear effects with true coefficients $\beta_{12} = 2$, $\beta_{13} = -0.2$. However the TRAM-DAG models the causal effect of X_2 on X_3 by a complex shift. In Fig. 17 (left side), we see the estimated complex shift term $-\hat{f}(X_2)$ for the case where the true function is $f(X_2) = -0.3 \cdot X_2$. We note small deviations of the fitted CS-term to the underlying linear function $f(x_2)$. As seen in the right panel of Fig. 17, both the observational distributions and interventional distributions show close alignment between the DGP and our trained model. The minor derivations have little effect on the quality of the estimated observational distributions, making training more challenging. Therefore we attribute the slight deviations to the limited training time and time data.

C.3.4. NON-MONOTONOUS DGP

To demonstrate that TRAM-DAGS are also able to fit complex non-monotonous functions the SCM, Fig. 18 presents the estimated coefficients β_{12} and β_{23}, along with the estimated non-monotonous function $\hat{f}(x_2)$ for $f(x_2) = 2\sin(3x_2) + x_2$.

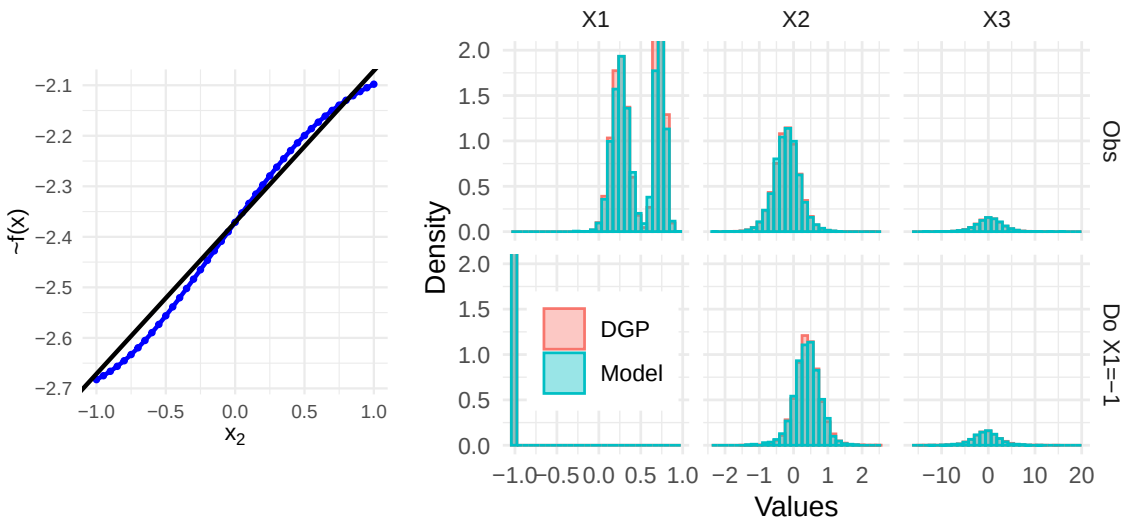

Figure 17: Interpretable continuous case experiment with three linear shift in DGP and two LS and one CS in the model: **Left:** Comparing the true linear effect $f(x_2) = 0.3 \cdot x_2$ of X_2 on X_3 in the DGP (black solid line) with the estimated CS $\hat{f}(X_2)$ in the fitted TRAM-DAG (blue dots). **Right:** Comparison of observational distributions (upper panel) and interventional distributions (lower panel) as generated by the DGP or the fitted TRAM-DAG.

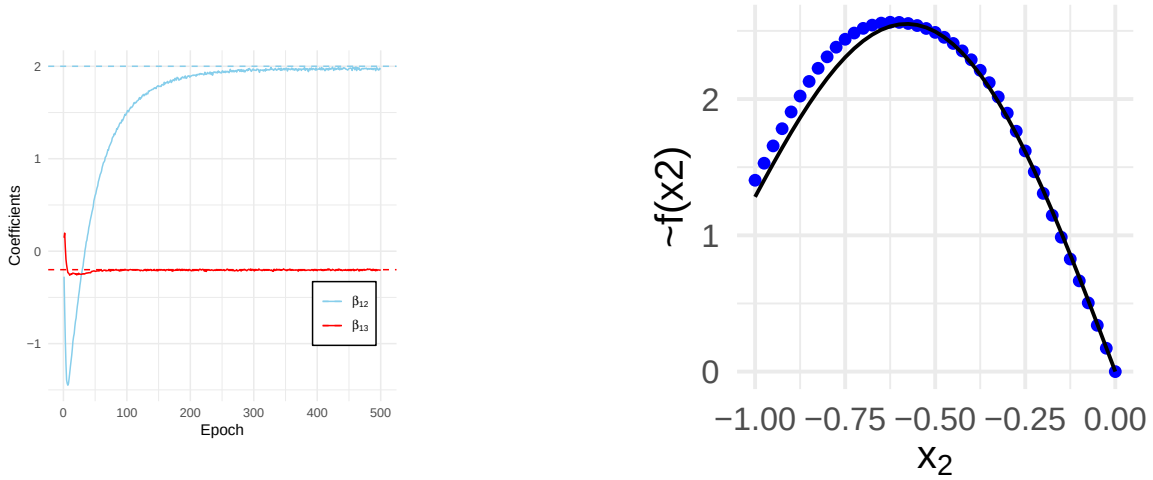

Figure 18: Interpretable continuous case experiment with two linear shift terms and one non-monotone shift terms in DGP and a correctly specified model: **Left:** Comparison of the true and estimated coefficients in the linear shift terms. **Right:** Comparison of the true non-monotone shift term $f(x_2) = 2\sin(3x_2) + x_2$ in the DGP (black solid line) and the fitted complex shift term in the TRAM-DAG (blue dots).

C.4. Interpretable Experiments Mixed Case

Here, we give additional results for the experiments for interpretable mixed TRAM-DAGs of Section 6.2 where X_3 is an ordinal variable (see Fig. 8).

Linear-shift DGP and linear-shift model We use in the DGP $f(X_2) = -0.3 \cdot X_2$ and fit a correctly specified TRAM-DAG model with linear shift terms for all variables. The estimated coefficients are in good agreement with the true values (see Fig. 19)..

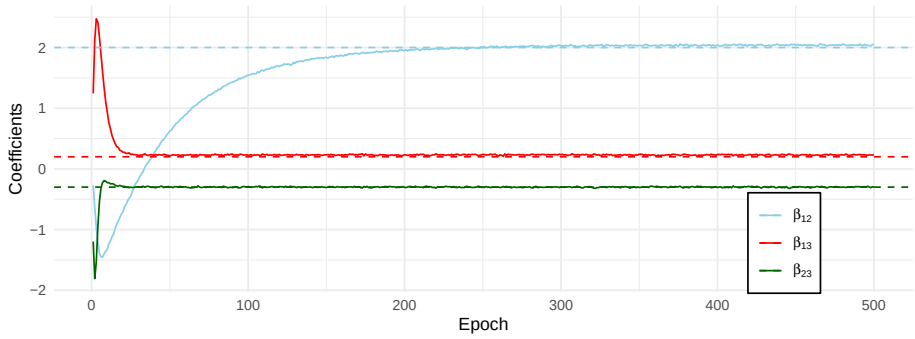

Figure 19: Interpretable experiments mixed case with only linear shift terms: Estimated coefficients β_{12}, β_{13}, and β_{23} over training epochs, with dashed lines representing the true coefficient values of the DGP.

Correctness of the predicted interventional effect To illustrate the interpretation of the estimated causal coefficients in the linear shift terms we use the transformation function $h(x_2|x_1) = SI + \beta_{12} \cdot x_1$ where in the DGP $\beta_{12} = 2$ which was estimated in the TRAM-DAG to be $\hat{\beta}_{12} = 2.05$.

Let's predict how the $\text{odds}(x_2 \leq c) = \frac{P(X_2 \leq c)}{1 - P(X_2 \leq c)}$ will change if x_1 is increased by one unit and choose $c = -1$.

According the theory of causal TRAM-DAGs the $\text{odds}(x_2 \leq c)$ should be in the interventional data changed by the factor of $e^{\hat{\beta}_{12}} = e^{2.05} = 7.74$ compared to this odds in the observational data. Hence the odds-ratio should be approximately $\hat{OR} = e^{2.05} = 7.74$.

To check this prediction we sample from the original DGP 40000 observations and then we adapt the DGP to a situation where X_1 is increased by one unit and sample also 40000 observation from the interventional distribution. Using this data we count in both samples how many x_2 observations were greater or not then the arbitray chosen cutoff $c = -1$ and receive the following number:

	number of x_2-values	
Type	≤ -1	> -1
Interventional	5119	34881
Observed	744	39256

This leads a point estimate of $\hat{OR} = 7.74$ and a 95% confidence interval $[7.16, 8.38]$ that holds the prediction of $\hat{OR} = e^{\hat{\beta}_{12}} = e^{2.05} = 7.74$ and the theoretical value of $OR = 7.4$. Hence, we were able to predict the correct change of the $\text{odds}(x_2 \leq -1)$ upon increasing $X1$ in the DGP by one unit from the estimated parameter $\hat{\beta}_{12} = 2.02$ of the TRAM-DAG that was fitted on observational data without having access to the interventional data.

Complex-shift DGP and complex-shift model Next, we increase the complexity of the DGP by defining $f(x_2) = 0.5 \cdot \exp(x_2)$ which introduces a non-linear causal impact of X_2 on X_3. We fit correctly specified TRAM-DAG and and show that the fitted model can be used to accurately estimate the observational and interventional distributions (see Fig. 20).

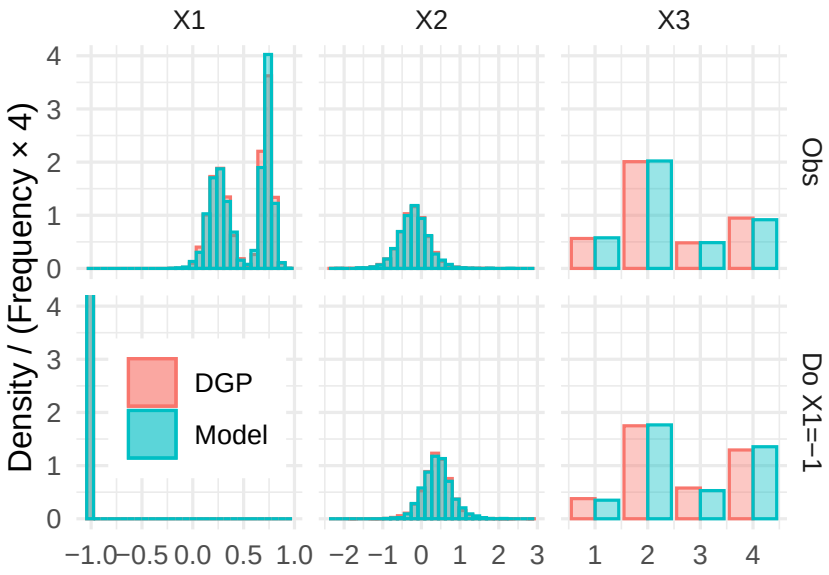

Figure 20: $\mathcal{L}_1$ and $\mathcal{L}_2$ for the mixed data experiment where the shift term from X_2 on X_3 were non-linear in the DGP with $f(x_2) = 0.5 \cdot \exp(x_2)$ and modeled as complex shift term in the fitted TRAM-DAG: Comparison of observational distributions (upper panel) and interventional distributions (lower panel) as generated by the mixed DGP or the fitted mixed TRAM-DAG. The frequencies of X_3 have been multiplied by a factor of 4 for visual convenience.

.