Learning Causal Structure on Mixed Data with Tree-Structured Functional Models

Tian Qin* Tian-Zuo Wang* Zhi-Hua Zhou*

Abstract

Discovering causal relations from observational data is at the heart of scientific research. Most causal discovery methods assume that the data have only one variable type. In real-world problems, however, data can consist of a mixture of continuous, discrete, and categorical variables. In this paper, we examine the causal discovery problem on mixed data. We introduce a general tree-structured functional causal model, which is well suited for characterizing the generating mechanisms of mixed data by allowing non-differentiability and nonlinearity. We present corresponding identifiability results, showing that under mild conditions, the causal directions can be uniquely determined from observational distributions. Further, we prove that the causal direction between continuous and discrete variables is generally identifiable under a much larger function class. Based on the theoretical findings, we propose an effective causal discovery method leveraging a consistent score function and powerful tree-learning techniques. Experiments on both synthetic and real data verify the effectiveness of our approach.

1 Introduction

Discovering causal relations among random variables is a fundamental and challenging task in science. Although interventions or randomized trials can be used for inferring causal relations, in many cases, they can be expensive, unethical, or even impossible. Thus, it is more desired to perform causal discovery from passively observed data [1, 2].

Causal relations are typically represented by a Directed Acyclic Graph (DAG) where nodes represent variables and directed edges depict direct causal relations. Under assumptions such as Markovness and faithfulness [2], constraint-based [1] and score-based [3] methods can identify partial causal relations from observational data and output a Markov equivalence class of DAGs, but still leave some causal directions undetermined. With the aid of Structural Equation Models (SEMs), the indeterminacy issue could be overcome under certain assumptions. Methods built upon SEMs [4, 5, 6] restrict the causal mechanisms to fall in specific function classes (e.g., linear non-Gaussian in Shimizu et al. [4]), then utilize the resultant asymmetries between the cause and effect variables to distinguish between DAGs in a Markov equivalence class. Though various function classes have been explored, most SEMs are built upon a common prerequisite: the variables share a common data type, e.g., variables are all continuous [4, 5, 6] or all discrete [7].

In real-world tasks, however, data often consist of a mixture of continuous (numerical), discrete (numerical), and categorical variables, thereby making most SEM-based methods inapplicable. Some beneficial efforts [8, 9, 10] have been made to design SEM-based methods to handle mixed data, but their applicability may be hindered by strict restrictions on the linearity of structural equations [8] or noise distributions [9, 10]. To make causal discovery more practical in real problems, it is necessary to investigate a new class of SEMs that is both capable of handling mixed data and general enough to approximate the underlying data generating process.

A central concern of designing an SEM-based method for mixed data is the choice of function class. Current SEMs are mostly from the differentiable function class, which excels in handling pure continuous data by leveraging the power of modern neural networks [6]. However, in mixed modeling tasks involving both categorical and numerical variables, the generating mechanisms can be nonlinear and non-differentiable. It is observed that tree-based models are more suited for such tasks [11]: trees can directly split on each possible value of a categorical variable without breaking the semantic meanings and utilize the power of ensembling [12]. Moreover, the modeling capabilities of trees can be further enhanced by incorporating neural networks as the predictive function in leaf nodes [13]. In this way, a hierarchical tree-structured neural network, which is at least as expressive as general neural networks even on pure numerical data, can be built. Thus, it can be beneficial to enrich the SEM family with flexible and powerful tree-structured models for handling mixed data.

In this paper, we study the causal discovery problem on mixed data under a class of general *tree-structured*

*National Key Laboratory for Novel Software Technology, Nanjing University, China, {qint,wangtz,zhouzh}@lamda.nju.edu.cn

SEMs. The concerned variables include continuous, discrete, and categorical ones, encompassing a wide range of real-world problems. We jointly model the generating process of various types of variables by allowing the causal mechanisms to be nonlinear and non-differentiable tree-structured functions. In contrast to previous works assuming the noise distribution function to have specific forms [9, 10], we only assume the noise variable to have continuous or discrete support. We also study the identifiability of the tree-structured SEM, i.e., finding under what conditions can one uniquely identify the causal directions from mixed data. We find that, under a large function class including the tree-structured ones, there is an intrinsic asymmetry between discrete and continuous variables that can benefit causal discovery, indicating that the differences in variable types may be a gift rather than an obstacle for identifying causal relations. Leveraging the asymmetry, we derive the bivariate identifiability results for possible combinations of the concerned variable types, showing that the causal directions between two variables are generally identifiable under mild conditions. Related results are extended to the multivariate case as well. Based on our theoretical findings, we develop an effective three-stage causal discovery method utilizing a consistent score function and powerful nonparametric tree-learning methods. To summarize, our contributions are mainly threefold:

1. We propose a tree-structured SEM for characterizing the causal mechanisms of mixed data.
2. We provide identifiability results for the proposed SEM, showing that under mild conditions, the causal directions can be uniquely determined from observational distributions. More general results on continuous-discrete causal relations under a larger function class are also presented.
3. We design a causal discovery method leveraging the tree-structured SEM. Experiments on both synthetic and real data verify its effectiveness.

2 Related Work

Constraint-based methods, such as PC [1], utilize conditional independence tests to recover the causal structure. Some proposals have been made to derive such tests for various variable types to adapt constraint-based methods for mixed data. Cui et al. [14] assumed that data were from a Gaussian copula model and built the independence tests upon the correlation matrix estimated by Gibbs sampling. Tsagris et al. [15] proposed to use likelihood ratio tests and derived symmetric conditional independence tests. Sedgewick et al. [16] applied a pseudo-likelihood method to learn an initial skeleton, then used the PC-Stable algorithm with likelihood ratio tests. Handhayani and Cussens [17] proposed a kernel alignment approach to computing a pseudo-correlation matrix that can be used in conditional independence test.

On the other hand, the score-based methods, such as GES [3], aim to find the causal structure by optimizing a score function. Efforts have been devoted to designing score functions for mixed data as well. Andrews et al. [18] proposed the conditional Gaussian score and mixed variable polynomial score. They assumed that given discrete variables, the continuous variables follow a multivariate Gaussian distribution. Later, Andrews et al. [19] derived a score function for data following a degenerate Gaussian distribution. Huang et al. [20] proposed a generalized score function that exploits regression in a reproducing kernel Hilbert space to capture the dependence between variables nonparametrically.

SEM-based methods for mixed data, which can avoid the indeterminacy issue of previously mentioned methods to some extent, have also been examined. Wei et al. [8] considered a class of linear additive noise functions for continuous and binary variables and assumed a Laplace or logistic distribution for the noise variable. Wei and Feng [9] assumed generating mechanisms to be three times differentiable nonlinear functions with additive Gaussian noise. Li et al. [10] assumed categorical variable distributions are specified by a softmax function by restricting the noise variables to follow Gumbel distributions.

3 Model Definition

We aim to recover the underlying causal DAG from m i.i.d. observed data points denoted by $D = \{(x_1^k, \dots, x_d^k)\}_{k=1}^m$. Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG with $\mathbf{V} = \{1, \dots, d\}$ denoting the node set and $\mathbf{E} \subseteq \mathbf{V}^2$ the edge set. A node j is a parent of i if $(j, i) \in \mathbf{E}$. The set of parents of i is denoted by $\mathbf{PA}_i$. We assume that each node i is associated with a random variable X_i which can be continuous (numerical), discrete (numerical) or categorical. If X_i is a discrete variable, we assume $X_i \in \mathbb{Z}$. If X_i is a categorical random variable with $T_i \geq 2$ categories, we assume $X_i \in [T_i] = \{1, \dots, T_i\}$.

As mentioned in Sec. 1, tree-structured functions are suited for describing mixed data. A decision tree typically divides the input space into disjoint subsets and makes predictions using functions defined on each subset. The function class where a decision tree reside is of the form $f(\mathbf{x}) = \sum_j \mathbb{I}_{\mathcal{R}_j}(\mathbf{x}) f_i(\mathbf{x})$, where $\{\mathcal{R}_j\}_j$ is a partition of the input space and $\mathbb{I}_{\mathcal{R}_j}(\mathbf{x})$ is the indicator function taking value 1 if $\mathbf{x} \in \mathcal{R}_j$ and 0 otherwise. More powerful tree ensemble models, such as random forest and gradient boosting decision trees, are typically a weighted average of trees, and thus can also be described by the above function. To make a flexible and expressive SEM, we mimic the tree structure by considering a partition of the parental space of each variable and propose the Tree-Structured Causal Model (TSCM).

Definition 3.1. (Parental partition) *Suppose that $X_i \in \mathcal{X}_i$ for $i \in [d]$. The* parental partition *of X_i is a set $\{\mathcal{R}_i^j\}_{j=1}^{K_i}$ such that $\bigcup_{j=1}^{K_i} \mathcal{R}_i^j = \prod_{u \in \mathbf{PA}_i} \mathcal{X}_u$ (Cartesian product) and $\bigcap_{j=1}^{K_i} \mathcal{R}_i^j = \emptyset$, where $K_i \geq 1$ is the size of the parental partition.*

Definition 3.2. (Tree-Structured Causal Model) *A* tree-structured causal model *is defined as a tuple $(\mathcal{S}, \mathcal{L}(\mathbf{N}))$, where $\mathcal{S} = (S_1, \dots, S_d)$ is a collection of d equations and $\mathcal{L}(\mathbf{N}) = \mathcal{L}(N_1, \dots, N_d)$ is the distribution of mutually independent noise variables $\{N_1, \dots, N_d\}$. Given a DAG $\mathcal{G}$, for each X_i, there is a parental partition $\{\mathcal{R}_i^j\}_{j=1}^{K_i}$ and an associated noise N_i independent of the parents $\mathbf{PA}_i$. If X_i is continuous (discrete), the generating mechanism is*

$$S_i: \quad X_i := \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}(\mathbf{PA}_i)\, f_i^j(\mathbf{PA}_i) + N_i, \tag{3.1}$$

where N_i is also continuous (discrete), $f_i^j : \prod_{u \in \mathbf{PA}_i} \mathcal{X}_u \to \mathbb{R}$ ($\mathbb{Z}$). If X_i is a categorical variable that can take T_i distinct values,

$$S_i: \quad X_i := \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}(\mathbf{PA}_i)\, f_i^j(N_i), \tag{3.2}$$

where N_i is continuous and $f_i^j : \mathbb{R} \to [T_i]$.

We assume that the observed data are generated by a TSCM. For continuous and discrete variables, we model the generating process with a region-based function and do not specify the function forms in each region, thus allowing for modeling non-differentiable and nonlinear relations. It is noteworthy that LiNGAM [4] and ANM [5] are special cases of Eq. (3.1) by setting K_i to 1, confirming the strong expressiveness of TSCMs. While for categorical variables, as they are not amenable to arithmetic operations such as addition, it may be unnatural to model an interaction between the parents and noise with certain functions. So we disentangle them and posit that the parents take effect only through the partition, and the noise decides how to generate a categorical value through $f_i^j(N_i)$.

The distribution $\mathcal{L}(\mathbf{X})$ generated by a TSCM is Markov w.r.t. the corresponding DAG $\mathcal{G}$ [2]. As a common practice [21], we further assume causal minimality, i.e., $\mathcal{L}(\mathbf{X})$ is Markov w.r.t. $\mathcal{G}$ but not to any proper subgraph of $\mathcal{G}$.

4 Identifiability

We next study the identifiability of the TSCM, which clarifies under what conditions can the causal structure be fully determined from observational data. The formal definition of identifiability is as follows:

Definition 4.1. (Identifiability) *Given a distribution law $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \dots, X_d)$ that has been generated by a TSCM with DAG $\mathcal{G}$, $\mathcal{G}$ is* identifiable *if $\mathcal{L}(\mathbf{X})$ cannot be generated by a TSCM with a DAG $\mathcal{G}' \neq \mathcal{G}$.*

As the identifiability issue between variables of a common type has been extensively studied [7, 21, 22], we focus on the identifiability of cases where variables have distinct types. Results on single-type cases can be found in Appendix. We first consider bivariate identifiability, then extend the results to multivariate cases. We start from the continuous-discrete case and consider a more general class of functions that contains TSCM as a special case. The following theorem shows that under such a function class, there exists an intrinsic asymmetry between cause and effect variables, which can greatly benefit causal discovery.

Theorem 4.1. *Suppose that $f(x, \cdot) : \mathbb{R} \to \mathbb{R}$ is invertible and continuous for $x \in \mathbb{Z}$, $g(y, \cdot) : \mathbb{Z} \to \mathbb{Z}$ is invertible for $y \in \mathbb{R}$, then a discrete random variable $X \in \mathbb{Z}$ and a continuous random variable $Y \in \mathbb{R}$, whose p.d.f. is continuous, can be described by at most one of Eq. (4.3) and Eq. (4.4) if $X \not\perp\!\!\!\perp Y$:*

$$Y = f(X, U) \quad \text{and} \quad X \perp\!\!\!\perp U, \tag{4.3}$$

$$X = g(Y, V) \quad \text{and} \quad Y \perp\!\!\!\perp V, \tag{4.4}$$

where $U \in \mathbb{R}$ and $V \in \mathbb{Z}$ are noise variables, and the p.d.f. of U is continuous and strictly positive over $\mathbb{R}$.

Remark 4.1. *As Eqs.* (4.3) *and* (4.4) *cannot hold simultaneously, Thm. 4.1 effectively states that the class of causal models that can be described by f and g is identifiable: $X \to Y$ if Eq.* (4.3) *holds and $X \leftarrow Y$ o.w. (it cannot be the case that both Eqs.* (4.3) *and* (4.4) *do not hold due to the causal minimality assumption). The functions f and g are general in the sense that they contain extensions of common SEMs to mixed data as special cases, including models with additive or multiplicative noise [4, 5, 7] and the post-nonlinear model [6]. We do not restrict the functions to be continuous or differentiable in their first arguments, which means that the tree-structured function in Eq.* (3.1) *belongs to this class as well. Moreover, the theorem is proved mainly by exploiting an* intrinsic *asymmetry between discrete and continuous variables, the difference between countable and uncountable sets, rather than relying on assumptions on specific functional forms. The intrinsic asymmetry can be seen as a hint telling us that identifying* *causal directions of a continuous-discrete pair may be intrinsically easier than in the single-type case.*

As a direct implication of Thm. 4.1, by assuming only a weak condition on the continuous part and without any additional restrictions on the discrete part, we have the identifiability for the continuous-discrete case in a TSCM:

Condition 4.1. *For a tuple $(\mathcal{L}(X_i), \mathcal{L}(N_i))$, where $X_i \in \mathbb{R}$, the p.d.f. of X_i is continuous, and the p.d.f. of N_i is continuous and strictly positive over $\mathbb{R}$.*

Corollary 4.1. *Suppose that a discrete random variable $X_1 \in \mathbb{Z}$ and a continuous random variable $X_2 \in \mathbb{R}$ are generated by a TSCM $((S_1, S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $(\mathcal{L}(X_2), \mathcal{L}(N_2))$ satisfies Cond. 4.1, then $\mathcal{G}$ is identifiable.*

For the continuous-categorical case, the categorical variable is generated by Eq. (3.2), which does not necessarily follow Eq. (4.4) since the noise can be real-valued and the function is typically not invertible in the second argument as required in Thm. 4.1. However, with analogous proof techniques, we find that the finite size of the parental partition in the TSCM, along with the uncountably many values the continuous variable can take, results in another asymmetry between the cause and effect variables. Therefore, Cond. 4.1 can also be applied to ensure the identifiability in this case:

Theorem 4.2. *Suppose that a categorical random variable $X_1 \in [T_1]$ and a continuous random variable $X_2 \in \mathbb{R}$ are generated by a TSCM $((S_1, S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $(\mathcal{L}(X_2), \mathcal{L}(N_2))$ satisfies Cond. 4.1, then $\mathcal{G}$ is identifiable.*

We next consider the discrete-categorical case. Unfortunately, the causal direction is not identifiable if no further constraints are imposed on a TSCM. We illustrate this point with the following proposition.

Proposition 4.1. *Given a joint observational distribution of a categorical random variable $X_1 \in [T_1]$ and a discrete random variable $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, by setting K_1 to $|\mathcal{X}_2|$ and each $\mathcal{R}_1^j$ to contain a unique element from $\mathcal{X}_2$ for $j \in [|\mathcal{X}_2|]$, there always exist functions $\{g_1^j\}_{j=1}^{K_1}$ and a random variable N_1' such that X_1 can be generated by*

$$X_1 := \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(X_2)\, g_1^j(N_1') \quad \textit{and} \quad X_2 \perp\!\!\!\perp N_1' \tag{4.5}$$

while giving the same observational distribution.

By comparing Eqs. (3.2) and (4.5), one can find that the causal direction from the discrete variable to the categorical one appears to hold no matter what the true causal direction is. An analogous result for the reverse direction holds as well. Therefore, no causal conclusions can be made. However, a necessary condition for Prop. 4.5 to hold is that the size of the partition of the parent variable is as large as the cardinality of the sample space, which is quite restrictive. In reality, it is common that some discrete values have identical effects on the generating process of another variable. For example, the age of a patient (discrete) directly causes which treatment strategy to use (categorical). It is rarely the case that every age value has a significant effect on the treatment choice. It is more reasonable to assume that the treatment choice is affected by age groups, forming a partition of a much smaller size. In light of this, it is reasonable to require the partition size to be smaller than the cardinality of the discrete sample space. The requirement and further conditions are stated in Cond. 4.2, leading to the identifiability results in Thm. 4.3.

Condition 4.2. *The tuple $((S_1, S_2), \mathcal{L}(X_1, X_2))$, where $X_1 \in \mathcal{X}_1 = [T_1]$ and $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, satisfies that if X_2 is the parent of X_1, the size of the parental partition $K_1 < |\mathcal{X}_2|$ and that if X_1 is the parent of X_2, the size of the parental partition $K_2 < |\mathcal{X}_1|$. Further, the following two conditions are not satisfied at the same time: (1) there exist two different values $a, b \in \mathcal{X}_2$, and a constant C, such that $\forall x \in \mathcal{X}_1$,*

$$\Pr(X_2 = a \mid X_1 = x) = C \cdot \Pr(X_2 = b \mid X_1 = x);$$

(2) there exist two different values $a, b \in \mathcal{X}_1$, and a constant C, such that $\forall x \in \mathcal{X}_2$,

$$\Pr(X_1 = a \mid X_2 = x) = C \cdot \Pr(X_1 = b \mid X_2 = x).$$

Theorem 4.3. *Suppose that a categorical random variable $X_1 \in [T_1]$ and a discrete random variable $X_2 \in \mathbb{Z}$ are generated by a TSCM $((S_1,\ S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $((S_1, S_2), \mathcal{L}(X_1, X_2))$ satisfies Cond. 4.2, then $\mathcal{G}$ is identifiable.*

Remark 4.2. *Cond. 4.2 requires that there does not exist a stable proportional relation between the conditional probabilities from two directions, which is reasonable to hold in real problems since a physical law ensuring the exact proportional relations in probability would be pathological and should be rare. This condition can also be used to give identifiability results in other scenarios. E.g., we find that it coincides with the assumption made to ensure the identifiability between two categorical variables in Cai et al. [22].*

Based on Peters et al. [21, Thm. 28], we extend the bivariate identifiability results to the multivariate case:

Theorem 4.4. *Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_d)$ be generated from a TSCM with graph $\mathcal{G}$. If $\mathcal{L}(\mathbf{X})$ satisfies causal minimality [2], and for all $j \in \mathbf{V}$, $i \in \mathbf{PA}_j$ and all sets $\mathbf{S} \subseteq \mathbf{V}$ with $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i, j\}$, where $\mathbf{ND}_j$ represents the non-descendents of j, there is an $x_{\mathbf{S}}$ with $p_{\mathbf{S}}\,(x_{\mathbf{S}}) > 0$ such that $\big(S_{i|x_{\mathbf{PA}_j\setminus\{i\}}},\, \mathcal{L}(X_i \mid X_{\mathbf{S}} = x_{\mathbf{S}}),\, \mathcal{L}(N_j)\big)$ satisfies the conditions for corresponding variable types in Cor. 4.1, Thm. 4.2-4.3, and Prop. B.1-B.3 (in Appendix, stating identifiability results for variables that share a common data type), then $\mathcal{G}$ is identifiable.*

5 Proposed Method

Assuming the conditions ensuring the identifiability (e.g., the conditions in the previous section) hold, there is a unique causal graph that can generate the observational data distribution, which means that it is possible to recover the underlying causal relations from purely observational data. A typical method based on SEMs is performing regression for each variable with the function family specified by the SEM, then measuring the independence between residuals (or noise) and the predictors [6, 21]. For example, in the bivariate case, if the independence only holds in one direction, we conclude that the direction is the causal one. However, in practice, the conditional independence tests are less reliable since we only have finite data points. In the mixed-data case, the problem is exacerbated since the independence tests must be performed between variables with distinct data types. Moreover, as shown in the following proposition, one can always find a noise variable that is dependent on the cause variables with a tree-structured function even from the correct causal direction in a TSCM, meaning that we cannot even approximately recover the true noise in Eq. (3.2).

Proposition 5.1. *For any categorical variable X_i generated from Eq. (3.2) with non-empty parents, there exist a random variable $N_i' \not\perp\!\!\!\perp \mathbf{PA}_i$ and functions $\{g_i^j\}_{j=1}^{K_i}$ such that generating X_i using*

$$X_i := \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}\left(\mathbf{PA}_i\right) g_i^j(N_i')$$

gives identical observational distribution of X_i and $\mathbf{PA}_i$.

As we cannot distinguish between the true noise N_i and another variable N_i' from observational data, independence measures between parents and noise in a TSCM are infeasible. Instead, we develop a causal discovery method by optimizing a consistent score function that distinguishes the true causal model from others in the sample limit [23] to avoid measuring independence.

Definition 5.1. (Penalized log-likelihood score) *Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and observed data $D=\{(x_1^k, \ldots, x_d^k)\}_{k=1}^m$, let $\mathcal{P}(\mathcal{G})$ denote the class of distributions that can be generated by $\mathcal{G}$. The* penalized log-likelihood score *is*

$$s\left(\mathcal{G}, D\right) = \max_{p \in \mathcal{P}(\mathcal{G})} \frac{1}{m} \sum_{k=1}^{m} \log p(x_1^k, \ldots, x_d^k) - \frac{|\mathbf{E}|}{\log m}.$$

Proposition 5.2. *Suppose that $\mathcal{G}^*$ is the identifiable true causal graph generating $D=\{(x_1^k, \ldots, x_d^k)\}_{k=1}^m$, under some technical conditions [23, Thm. 1], the penalized log-likelihood score is consistent, i.e., $\forall\ \mathcal{G} \neq \mathcal{G}^*$,*

$$\Pr\left(s\left(\mathcal{G}, D\right) < s\left(\mathcal{G}^*, D\right)\right) \to 1 \quad as \quad m \to \infty.$$

The penalized log-likelihood score consists of a maximum log-likelihood and a regularization term penalizing graphs with more edges. Prop. 5.2 shows that the score is consistent in the sense that the true causal graph attains the maximum score asymptotically. Assuming that the causal graph is identifiable under a TSCM, the causal discovery problem turns into finding the DAG with a maximized score.

The question arises of how to compute the log-likelihood score for a TSCM. Since the tree-structured functions are generally fitted non-parametrically and we do not specify the noise distribution, the likelihood is not directly available. As the distribution is Markov with respect to the causal DAG, we factorize the probability of the observed data into

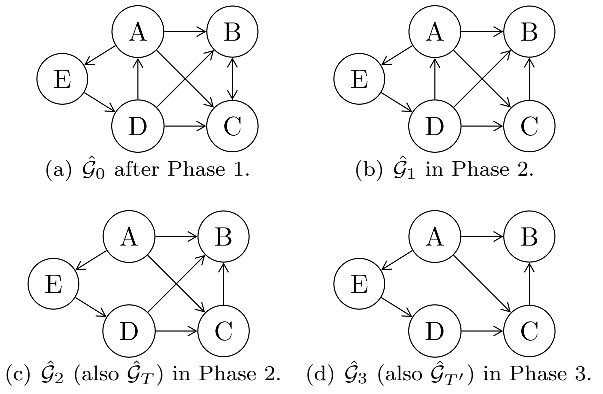
$$(5.6) \qquad \prod_{k=1}^{m} p\left(x_1^k, \ldots, x_d^k\right) = \prod_{k=1}^{m} \prod_{i=1}^{d} p\left(x_i^k \mid \mathbf{pa}_i^k\right).$$

For a categorical X_i, by fitting a tree-based classifier over its parents, we obtain a posterior probability of X_i which is an estimate for $p(x_i^k \mid \mathbf{pa}_i^k)$. For a numerical X_i, according to Eq. (3.1), we can rewrite the conditional probability with

$$p\left(x_i^k \mid \mathbf{pa}_i^k\right) = p_{n_i}\left(x_i^k - \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}\left(\mathbf{pa}_i^k\right) f_i^j\left(\mathbf{pa}_i^k\right)\right)$$

where p_{n_i} is the probability density (mass) function of the noise N_i if X_i is continuous (discrete). By regressing each numerical variable on its parents with a tree-based regressor such as a random forest, we obtain an approximation $\hat{f}_i$ for the region-based function and a noise estimate $\hat{n}_i = x_i^k - \hat{f}_i(\mathbf{pa}_i^k)$. If the noise is continuous, we estimate its density with kernel density estimation [24]. If the noise is discrete, an empirical probability mass function is available by counting the frequencies of each value. Plugging the estimates into Eq. (5.6) gives us a computable likelihood function.

In order to find a DAG maximizing the penalized log-likelihood score, a direct method is to enumerate every possible DAG and select the one with the maximum score. The search space grows super-exponentially with d [3], meaning that brute-force search is only applicable to cases with a very small number of variables. Referring to Bühlmann et al. [25], we propose a local search method named **TR**ee-b**A**sed **C**ausal discovery by greedy **E**dge **R**emoving (TRACER) to find the DAG with the maximum score. TRACER is divided into three phases. We first apply *variable selection* in Phase 1 to reduce the search space. Then in Phase 2, starting from an initial overly connected directed graph, we greedily remove redundant edges that give a higher score until we have a valid DAG. In Phase 3, we again run variable selection as a post-processing step to make the found structure sparse and further increase the score. An illustration of TRACER is given in Fig. 1 and details of the three phases are given as follows:



(a) $\hat{\mathcal{G}}_0$ after Phase 1. (b) $\hat{\mathcal{G}}_1$ in Phase 2.

(c) $\hat{\mathcal{G}}_2$ (also $\hat{\mathcal{G}}_T$) in Phase 2. (d) $\hat{\mathcal{G}}_3$ (also $\hat{\mathcal{G}}_{T'}$) in Phase 3.

Figure 1: An illustration of the TRACER algorithm. The initial graph (a) after Phase 1 contains directed cycles $A \to E \to D \to A$ and redundant edges $B \to C$ and $B \leftarrow C$ (which also constitutes a length-2 directed cycle). Then in Phase 2, edges are removed in (b) and (c) greedily according to the gain in score until we have a valid DAG $\hat{\mathcal{G}}_T$. In Phase 3, redundant edges are further pruned to increase the overall score and we get a final output in (d).

Phase 1. *Search space reduction by variable selection.* For each numerical (categorical) variable X_i, perform regression (classification) using X_i as the target variable and $\{X_j\}_{j \neq i}$ as predictors with a variable selection method. We use the optimal decision tree [26, 27] as the variable selection method to filter out variables that are not very predictive for a target variable as the generating mechanisms considered in this paper have a tree structure, but other selection methods may be applied as well. The selected predictors form a candidate parental set $\widehat{\mathbf{PA}}_i$ for X_i.

Phase 2. *DAG search by greedy edge removing.* We start from a graph $\hat{\mathcal{G}}_0 = (\mathbf{V}, \mathbf{E}_0)$ containing all edges pointing from variables in the candidate parent sets $\widehat{\mathbf{PA}}_i$ to corresponding X_i. $\hat{\mathcal{G}}_0$ is a directed graph possibly with directed cycles and bi-directed edges. The following procedure is repeated until we get a valid DAG $\hat{\mathcal{G}}_T$: in iteration $t \geq 1$, we obtain a new graph $\hat{\mathcal{G}}_t = (\mathbf{V}, \mathbf{E}_t)$ by setting $\mathbf{E}_t$ to $\mathbf{E}_{t-1} \setminus \{(u, v)\}$, where (u, v) is an edge that is part of a directed cycle in $\hat{\mathcal{G}}_{t-1}$ and by removing it we get a locally maximized score. The procedure of finding the edge to remove is in Alg. 1.

Phase 3. *DAG refinement by variable selection.* We run variable selection for each variable on corresponding parents encoded by $\hat{\mathcal{G}}_T$ to further refine the graph structure. Let $\mathbf{\Gamma}_i$ denote the parents of X_i that are *not* selected by the variable selection method, and $\mathbf{C} \triangleq \bigcup_{i \in [d]} \{(j, i) \mid j \in \mathbf{\Gamma}_i\}$. We repeat the following procedure: in each iteration, we remove an edge $(j, i) \in \mathbf{C}$ that gives a locally maximized score. The procedure stops until it reaches a local maxima $s(\hat{\mathcal{G}}_{T'}, D)$. Finally, $\hat{\mathcal{G}}_{T'}$ is output as the found causal structure.

Algorithm 1 DAG search by greedy edge removing

```
Input: $(\mathbf{V}, \mathbf{E}_0)$
for $i \in \mathbf{V}$ do
    $s_i \leftarrow \frac{1}{m}\sum_{k=1}^{m} \log \hat{p}(x_i^k \mid \hat{\mathbf{pa}}_i^k)$
$t \leftarrow 0$, $\delta \leftarrow -\infty$
while $(\mathbf{V}, \mathbf{E}_t)$ is not a DAG do
    for $(j, i) \in \mathbf{E}_t$ do
        if there is a directed path from $i$ to $j$ in $\mathbf{E}_t$ then
            $s_i' \leftarrow \frac{1}{m}\sum_{k=1}^{m} \log \hat{p}(x_i^k \mid \hat{\mathbf{pa}}_i^k \setminus \{x_j\})$
            if $s_i' - s_i > \delta$ then
                $\delta \leftarrow s_i' - s_i$, $(u, v) \leftarrow (j, i)$
    $\mathbf{E}_{t+1} \leftarrow \mathbf{E}_t \setminus \{(u, v)\}$, $\widehat{\mathbf{PA}}_v \leftarrow \widehat{\mathbf{PA}}_v \setminus \{X_u\}$
    $s_v \leftarrow s_v'$, $t \leftarrow t + 1$, $\delta \leftarrow -\infty$
$\hat{\mathcal{G}}_T \leftarrow (\mathbf{V}, \mathbf{E}_t)$
Output: $\hat{\mathcal{G}}_T$
```

The reason that we only remove edges in the proposed method is that removing edges is more likely to output sparse graphs than dense ones, which can be less useful than sparse structures in real applications, though the true causal graph may not be sparse at all. Developing methods that combine the proposed removing process and an edge-adding process, and still yield sparse graphs could be interesting and is left for future work. The running time of TRACER mainly depends on the number of calls to the log-likelihood estimation procedure, which first performs tree-based learning and then estimates probability density. The number of iterations in Phases 2 and 3 is at most $O(d^2)$ since an edge is removed in each iteration and there are $O(d^2)$ edges. The iterations can be much fewer if the variable selection method has filtered most superfluous edges in Phase 1. The log-likelihood gain associated with each edge removal can be calculated incrementally. Specifically, in Phase 2, suppose we are at iteration $t \geq 1$ and the edge removed in iteration $t-1$ is (j, i). As the log-likelihood can be factorized using Eq. (5.6) and the gain of removing edges into variables other than X_i has been computed in previous iterations, we can save them and skip line 7 in Alg. 1 for such edges, and only compute $\sum_{k=1}^{m} \log p(x_i^k \mid \mathbf{pa}_i^k)$ for $O(d)$ different $\mathbf{pa}_i^k$. So the overall number of calls to the log-likelihood estimation procedure can be reduced to $O(d^3)$.

6 Experiments

We apply the proposed TRACER method to both synthetic and real-world datasets to verify its effectiveness.

Implementations. We use optimal decision trees [26], a tree learning method that can learn a sparse and accurate tree, as the variable selection method in Phases 1 and 3. Most tree-based learning methods can be used in Phase 2. Here we apply random forest as the basic learning procedure for its efficiency and wide applicability. When the target variable is discrete, we treat it as a continuous one when fitting a regression model. A post-processing step is made to replace the predicted value with the nearest one from the discrete sample space. We used a kernel density estimator with an RBF kernel to estimate the noise density. The kernel width is set to twice the median distance between input points, as suggested by Huang et al. [20]. Note that when the causal discovery task only involves two variables, we skipped Phases 1 and 3 as variable selection would be effectless.

Baselines and metrics. For multivariate cases, we compare TRACER with constraint-based methods for mixed data: Copula PC [14] and CausalMGM [16], score-based methods for mixed data: Degenerate Gaussian score (DG) [19] and the Generalized Score (GS) [20] with greedy equivalence search [3]. Since these methods output an equivalence class, we further compare with methods that output a unique DAG, including DirectLiNGAM [28], CAM [25] and NOTEARS [29]. For bivariate cases, in addition to those that have been mentioned, we also compared the Additive Noise Model (ANM) [5], the Post-NonLinear model (PNL) [6] and Information Geometric Causal Inference (IGCI) [30]. We use the F1 score, Structural Hamming Distance (SHD), and Structural Intervention Distance (SID) as the evaluation metrics for multivariate cases. We normalize SHD and SID by dividing the number of edges and $d(d-1)$ respectively. For methods that output a Markov equivalence class, we provide the lower and upper bounds attainable by members within the equivalence class or report the average of the two bounds. For bivariate cases, we report the accuracy.

6.1 Synthetic Data

Data Generation. To generate datasets from TSCMs, we specify the number of variables d, the number of variables of each type, and the number of edges e. In some experiments, we specify the graph density, defined as $\rho \triangleq 2e/d^2$, instead of directly specifying e. We randomly set the data type of each variable to be continuous, discrete, or categorical to satisfy the requirement. An undirected graph is sampled from the Erdős-Rényi model with the expected number of edges equal to e. We obtain a DAG by orienting the undirected edges following a uniformly randomly picked ordering of the d variables. Following the topological order in the DAG, we generate the value of each variable from the value of its parents and corresponding noise. Due to limited space, we put more details in Appendix.

We conduct experiments on synthetic data generated from a TSCM where the functions in each region of the parental partitions are parameterized by a linear function or a randomly initialized neural network. For each combination of the number of variables d, expected number of edges e and the number of variables having different types, we report results over 100 realizations. We first make a simple test for TRACER on graphs with $d = 10$ nodes using $m = 1000$ samples in each realization. Fig. 2 shows that TRACER excels at handling mixed data and is robust against all three metrics: TRACER outperforms baselines on SHD and F1 scores and is comparable to the best baseline in terms of SID. The reason why most baselines do not perform well is probably that they make restrictive assumptions on the data generating process, but the assumptions (e.g., data are from a Gaussian copula model in Cui et al. [14]) are severely violated in this setting. GS does not assume a specific functional form, thus achieving a small SID.

Sensitivity to the proportion of categorical variables. We run experiments with the proportion of categorical variables ranging from 0.1 to 0.5 and $d \in \{10, 20, 30\}$. The graph density is set to 0.2. Fig. 3 shows that as the proportion of categorical variables increases, the performance of baselines decreases, whereas TRACER steadily outperforms baselines. This phe-

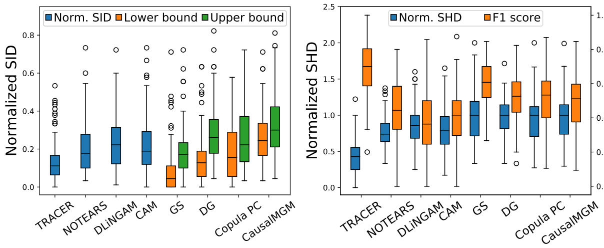


Figure 2: Box plots of F1 score, normalized SHD and SID with $d = e = 10$ and number of continuous/discrete/categorical variables 4/3/3. For methods that output an equivalence class, the upper and lower bound of the normalized SID are plotted.

Dataset	TRACER	CAM	NOTEARS	ANM	PNL	IGCI
CE pairs	63.7%	51.3%	54.4%	54.9%	57.4%	57.9%
Abalone	92.9%	42.9%	42.9%	71.4%	50.0%	100%

Table 1: Accuracy on CE pairs and Abalone dataset.

nomenon suggests that TRACER is more suited for handling mixed data, especially on data with a large proportion of categorical variables, by inheriting the merits of tree-structured models.

Sensitivity to dataset size, graph densities, and the number of variables. We conduct experiments with varying m, d, and ρ. Fig. 4 shows that TRACER achieves superior performance on all three metrics in most settings, which further verifies its effectiveness. Although the GS method achieves a comparable SID in some cases, it is computationally too costly to scale to large graphs. For $d = 50$, we could not finish experiments on GS within 48 hours.

6.2 Real Data The Cause-Effect (CE) pairs challenge dataset[1] consists of real and semi-artificial variable pairs. We run experiments on 594 numerical-categorical pairs that have causal relations. Tbl. 1 shows that TRACER outperforms other baselines by a large margin. We also conduct experiments on the Abalone dataset[2] which contains the age (discrete), sex (categorical), and some real-valued physical measurements of abalone. Both age and sex cause other variables. TRACER successfully identifies 13 out of 14 causal relations. The results in Tbl. 1 indicate that on real data that are not necessarily generated from a TSCM, TRACER is still very effective.

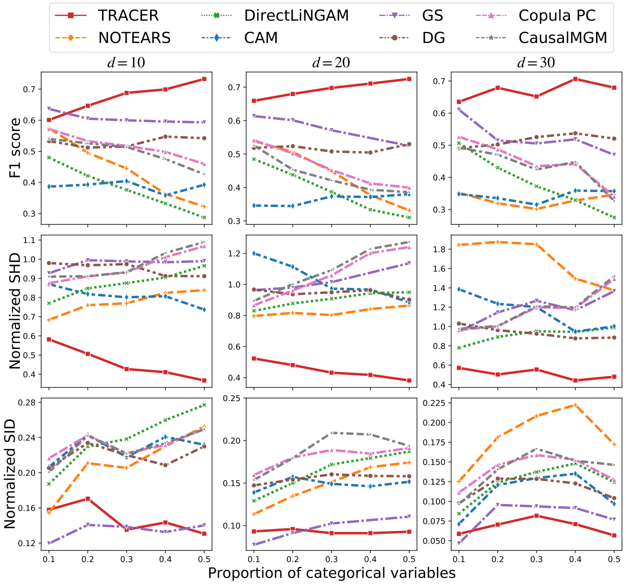


Figure 3: Averaged F1 score, normalized SHD and SID over 100 simulations with varying proportions of categorical variables. The proportion of discrete variables is set to 0.3. For methods that output a Markov equivalence class, the SID curve depicts the average of upper and lower bound for better readability.

7 Conclusion

In this paper, we examine causal discovery on mixed observational data that contain continuous, discrete, and categorical variables. We introduce a flexible and expressive tree-structured causal model that allows non-differentiability and non-linearity. We theoretically analyze its identifiability and propose an effective tree-based causal discovery method. Experiments on both synthetic and real data verify the superiority of the proposed method.

Acknowledgement

This research was supported by the National Science Foundation of China (61921006). The authors thank Jia-Wei Shan and Qin-Cheng Zheng for the discussion and thank the anonymous reviewers for their comments.

References

[1] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2000.

[2] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[3] D. M. Chickering, “Optimal structure identification with greedy search,” *JMLR*, vol. 3, pp. 507–554, 2002.

[4] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, “A linear non-Gaussian acyclic model for

[1] http://www.causality.inf.ethz.ch/cause-effect.php

[2] https://archive.ics.uci.edu/ml/datasets/abalone

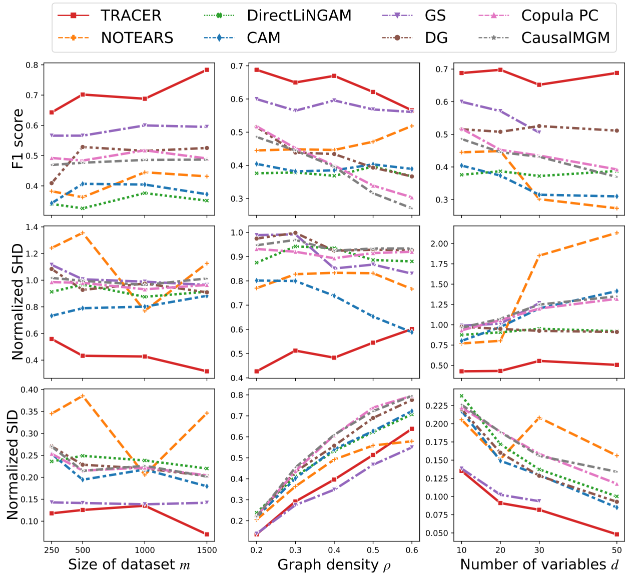


Figure 4: Averaged F1 score, normalized SHD and SID over 100 simulations with varying $m/\rho/d$. The proportion of continuous/discrete/categorical variables is 0.4/0.3/0.3 respectively. Unless specified by the x-axis, we set $m = 1000$, $d = 10$, and $\rho = 0.2$.

causal discovery," *JMLR*, vol. 7, pp. 2003–2030, 2006.

[5] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *NeurIPS*, 2008, pp. 689–696.

[6] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *UAI*, 2009, pp. 647–655.

[7] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on discrete data using additive noise models," *TPAMI*, vol. 33, no. 12, pp. 2436–2450, 2011.

[8] W. Wei, L. Feng, and C. Liu, "Mixed causal structure discovery with application to prescriptive pricing," in *IJCAI*, 2018, pp. 5126–5134.

[9] W. Wei and L. Feng, "Nonlinear causal structure learning for mixed data," in *ICDM*, 2021, pp. 709–718.

[10] Y. Li, R. Xia, C. Liu, and L. Sun, "A hybrid causal structure learning algorithm for mixed-type data," in *AAAI*, 2022, pp. 7435–7443.

[11] Z.-H. Zhou and J. Feng, "Deep Forest," *Natl. Sci. Rev.*, vol. 6, no. 1, pp. 74–86, 10 2018.

[12] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012.

[13] Z.-H. Zhou and Z. Chen, "Hybrid decision tree," *Knowl. Based Syst.*, vol. 15, no. 8, pp. 515–528, 2002.

[14] R. Cui, P. Groot, and T. Heskes, "Copula PC algorithm for causal discovery from mixed data," in *ECML PKDD*, 2016, pp. 377–392.

[15] M. Tsagris, G. Borboudakis, V. Lagani, and I. Tsamardinos, "Constraint-based causal discovery with mixed data," *Int. J. Data Sci. Anal.*, vol. 6, no. 1, pp. 19–30, 2018.

[16] A. J. Sedgewick, K. Buschur, I. Shi, J. D. Ramsey, V. K. Raghu, D. V. Manatakis, Y. Zhang, J. Bon, D. Chandra, C. Karoleski, F. C. Sciurba, P. Spirtes, C. Glymour, and P. V. Benos, "Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis," *Bioinform.*, vol. 35, no. 7, pp. 1204–1212, 2019.

[17] T. Handhayani and J. Cussens, "Kernel-based approach for learning causal graphs from mixed data," in *PGM*, 2020, pp. 221–232.

[18] B. Andrews, J. D. Ramsey, and G. F. Cooper, "Scoring Bayesian networks of mixed variables," *Int. J. Data Sci. Anal.*, vol. 6, no. 1, pp. 3–18, 2018.

[19] ——, "Learning high-dimensional directed acyclic graphs with mixed data-types," in *KDD Workshop on Causal Discovery*, 2019, pp. 4–21.

[20] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *KDD*, 2018, pp. 1551–1560.

[21] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *JMLR*, vol. 15, no. 1, pp. 2009–2053, 2014.

[22] R. Cai, J. Qiao, K. Zhang, Z. Zhang, and Z. Hao, "Causal discovery from discrete data using hidden compact representation," in *NeurIPS*, 2018, pp. 2671–2679.

[23] C. Nowzohour and P. Bühlmann, "Score-based causal learning in additive noise models," *Stat.*, vol. 50, no. 3, pp. 471–485, 2016.

[24] A. Z. Zambom and D. Ronaldo, "A review of kernel density estimation with applications to econometrics," *Int. Econ. Rev.*, vol. 5, no. 1, pp. 20–42, 2013.

[25] P. Bühlmann, J. Peters, and J. Ernest, "CAM: Causal additive models, high-dimensional order search and penalized regression," *Ann. Stat.*, vol. 42, no. 6, pp. 2526–2556, 2014.

[26] D. Bertsimas and J. Dunn, "Optimal classification trees," *Mach. Learn.*, vol. 106, no. 7, pp. 1039–1082, 2017.

[27] ——, *Machine Learning under a Modern Optimization Lens*. Dynamic Ideas LLC, 2019.

[28] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *JMLR*, vol. 12, pp. 1225–1248, 2011.

[29] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: continuous optimization for structure learning," in *NeurIPS*, 2018, pp. 9492–9503.

[30] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf, "Information-geometric approach to inferring causal directions," *Artif. Intell.*, vol. 182-183, pp. 1–31, 2012.

A Proofs

THEOREM 4.1. *Suppose that $f(x,\cdot):\mathbb{R}\to\mathbb{R}$ is invertible and continuous for $x\in\mathbb{Z}$, $g(y,\cdot):\mathbb{Z}\to\mathbb{Z}$ is invertible for $y\in\mathbb{R}$, then a discrete random variable $X\in\mathbb{Z}$ and a continuous random variable $Y\in\mathbb{R}$, whose p.d.f. is continuous, can be described by at most one of Eq. (4.3) and Eq. (4.4) if $X \not\perp\!\!\!\perp Y$:*

$$Y = f(X,U) \quad \text{and} \quad X \perp\!\!\!\perp U, \tag{4.3}$$

$$X = g(Y,V) \quad \text{and} \quad Y \perp\!\!\!\perp V, \tag{4.4}$$

where $U\in\mathbb{R}$ and $V\in\mathbb{Z}$ are noise variables, and the p.d.f. of U is continuous and strictly positive over $\mathbb{R}$.

Proof. Let $f_x(\cdot)\triangleq f(x,\cdot)$ and $g_y(\cdot)\triangleq g(y,\cdot)$. If both models hold, then

$$\Pr(X=x)p_u\left(f_x^{-1}(y)\right) = p(x,y) = p_y(y)\Pr\left(V=g_y^{-1}(x)\right). \tag{A.1}$$

Given x, since $f_x(\cdot)$ is surjective, we have $p(x,y)>0$ for any y as long as $\Pr(X=x)>0$. For a fixed x_0 with $\Pr(X=x_0)>0$, we can write

$$\frac{\Pr(X=x_0)}{\Pr\left(V=g_y^{-1}(x_0)\right)} = \frac{p_y(y)}{p_u\left(f_{x_0}^{-1}(y)\right)}. \tag{A.2}$$

Let $A\triangleq\left\{\frac{\Pr(X=x_0)}{\Pr\left(V=g_y^{-1}(x_0)\right)} \mid y\in\mathbb{R}\right\}$ and $B\triangleq\left\{\frac{p_y(y)}{p_u\left(f_{x_0}^{-1}(y)\right)} \mid y\in\mathbb{R}\right\}$. We next prove by contradiction. Suppose that there exists two different $y_1,y_2\in B$ such that $y_1<y_2$. Due to the continuity of $p_y(\cdot)$, $f_{x_0}^{-1}(\cdot)$ and $p_u(\cdot)$ and the positivity of $p_u(\cdot)$, we have $h(y)=\frac{p_y(y)}{p_u\left(f_{x_0}^{-1}(y)\right)}$ is also continuous. Using the intermediate value theorem, we have $[y_1,y_2]\subseteq B$, so $|B|\ge|\mathbb{R}|$. On the other hand, $|A|\le|\mathbb{Z}|$ since V can only take at most countably many values. Consequently, we have $|A|\le|\mathbb{Z}|<|\mathbb{R}|\le|B|$, which is a contradiction since $|A|=|B|$ according to Eq. (A.2). Thus, B contains only one element and $|B|=|A|=1$, meaning that

$$\frac{\Pr(X=x_0)}{\Pr\left(V=g_y^{-1}(x_0)\right)} = C, \tag{A.3}$$

where C is a constant. We have

$$p(y\mid x_0) = \frac{p(x_0,y)}{\Pr(X=x_0)} \tag{A.4}$$

$$= \frac{p_y(y)\Pr\left(V=g_y^{-1}(x_0)\right)}{\Pr(X=x_0)} \tag{A.5}$$

$$= \frac{p_y(y)}{C}. \tag{A.6}$$

Integrating both sides, we get

$$C = 1. \tag{A.7}$$

Thus,

$$p(y\mid x_0) = p_y(y).$$

Since the choice of x_0 was arbitrary, we have $X\perp\!\!\!\perp Y$, which contradicts the condition that $X\not\perp\!\!\!\perp Y$. So we conclude that at most one of Eq. (4.3) and Eq. (4.4) holds.

□

CONDITION 4.1. *For a tuple $(\mathcal{L}(X_i),\mathcal{L}(N_i))$, where $X_i\in\mathbb{R}$, the p.d.f. of X_i is continuous, and the p.d.f. of N_i is continuous and strictly positive over $\mathbb{R}$.*

COROLLARY 4.1. *Suppose that a discrete random variable $X_1\in\mathbb{Z}$ and a continuous random variable $X_2\in\mathbb{R}$ are generated by a TSCM $((S_1,S_2),\ \mathcal{L}(N_1,N_2))$ with a graph $\mathcal{G}:X_1\to X_2$ or $\mathcal{G}:X_1\leftarrow X_2$. If $(\mathcal{L}(X_2),\mathcal{L}(N_2))$ satisfies Cond. 4.1, then $\mathcal{G}$ is identifiable.*

Proof. The conclusion follows directly from Thm. 4.1 by noting that the tree-structured function S_2 with additive noise is invertible and continuous in the noise variable N_2 and that S_1 is invertible in N_1. □

THEOREM 4.2. *Suppose that a categorical random variable $X_1 \in [T_1]$ and a continuous random variable $X_2 \in \mathbb{R}$ are generated by a TSCM $((S_1, S_2), \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $(\mathcal{L}(X_2), \mathcal{L}(N_2))$ satisfies Cond. 4.1, then $\mathcal{G}$ is identifiable.*

Proof. The proof is analogous to that of Thm. 4.1. Inheriting the symbols used in Thm. 4.1, we use $f(x_1, n_2)$ to represent the assigned tree-structured function in S_2 (see Eq. (3.1)) and use $g(x_2, n_1)$ to represent the assigned one in S_1 (see Eq. (3.2)). We aim to prove that at most one of Eq. (A.8) and (A.9) can hold:

$$X_1 = g(X_2, N_1) = \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(X_2)\, f_1^j(N_1) \quad \text{and} \quad X_2 \perp\!\!\!\perp N_1, \tag{A.8}$$

$$X_2 = f(X_1, N_2) = \sum_{j=1}^{K_2} \mathbb{I}_{\mathcal{R}_2^j}(X_1)\, f_2^j(X_1) + N_2 \quad \text{and} \quad X_1 \perp\!\!\!\perp N_2. \tag{A.9}$$

Note that although f still satisfies the condition in Thm. 4.1, g is not. So we cannot directly apply Thm. 4.1.

From Eq. (A.8), we can write

$$\Pr(X_1 = x_1 \mid X_2) = \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(X_2)\, h_j(x_1), \tag{A.10}$$

where $h_j(x_1) = \Pr\left(f_1^j(N_1) = x_1\right)$. Let $f_{x_1}(\cdot) \triangleq f(x_1, \cdot)$, if Eq. (A.8) and (A.9) both hold, then

$$\Pr(X_1 = x_1) p_{n_2}\left(f_{x_1}^{-1}(x_2)\right) = p(x_1, x_2) = p(x_2) \Pr(X_1 = x_1 \mid x_2) = p(x_2) \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(x_2)\, h_j(x_1). \tag{A.11}$$

Given x_1, since $f_{x_1}(\cdot)$ is surjective, we have $p(x_1, x_2) > 0$ for any x_2 as long as $\Pr(X_1 = x_1) > 0$. For a fixed x_1 with $\Pr(X_1 = x_1) > 0$, we can write

$$\frac{\Pr(X_1 = x_1)}{\sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(x_2)\, h_j(x_1)} = \frac{p(x_2)}{p_{n_2}\left(f_{x_1}^{-1}(x_2)\right)}. \tag{A.12}$$

The proof left is then analogous to that of Thm. 4.1 by noting that $\sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(x_2)\, h_j(x_1)$ can take at most finite K_1 values for a fixed x_1, which again leads to a contradiction of set cardinalities if the above equation can have more than one values. ◻

PROPOSITION 4.1. *Given a joint observational distribution of a categorical random variable $X_1 \in [T_1]$ and a discrete random variable $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, by setting K_1 to $|\mathcal{X}_2|$ and each $\mathcal{R}_1^j$ to contain a unique element from $\mathcal{X}_2$ for $j \in [|\mathcal{X}_2|]$, there always exist functions $\{g_1^j\}_{j=1}^{K_1}$ and a random variable N_1' such that X_1 can be generated by*

$$X_1 := \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(X_2)\, g_1^j(N_1') \quad \text{and} \quad X_2 \perp\!\!\!\perp N_1' \tag{4.5}$$

while giving the same observational distribution.

Proof. The joint distribution of X_1 and X_2 can be factorized as

$$p(x_1, x_2) = p(x_2) p(x_1 \mid x_2). \tag{A.13}$$

Let $N_1' \sim \text{Uniform}(0, 1)$ be a random variable independent of X_2. For a fixed $x_2 \in \mathbb{Z}$, let $\alpha_i \triangleq p(X_1 = i \mid x_2)$, $i \in [T_1]$, and $g_1^{x_2}(N_1') = \sum_{k=1}^{T_1} \mathbb{I}\left(N_1' \in \left[\sum_{i=0}^{k-1} \alpha_i, \sum_{i=0}^{k} \alpha_i\right]\right) k$. Let a new random variable X_1' be generated from Eq. (4.5), we have the conditional probability

$$p(X_1' = i \mid x_2) = \alpha_i = p(X_1 = i \mid x_2), \tag{A.14}$$

which gives

$$p(x_1, x_2) = p(x_1', x_2), \tag{A.15}$$

so X_1 can be generated from Eq. (4.5) and we have the same observational distribution. ◻

Condition 4.2. *The tuple $((S_1, S_2), \mathcal{L}(X_1, X_2))$, where $X_1 \in \mathcal{X}_1 = [T_1]$ and $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, satisfies that if X_2 is the parent of X_1, the size of the parental partition $K_1 < |\mathcal{X}_2|$ and that if X_1 is the parent of X_2, the size of the parental partition $K_2 < |\mathcal{X}_1|$. Further, the following two conditions are not satisfied at the same time: (1) there exist two different values $a, b \in \mathcal{X}_2$, and a constant C, such that $\forall x \in \mathcal{X}_1$,*

$$\Pr(X_2 = a \mid X_1 = x) = C \cdot \Pr(X_2 = b \mid X_1 = x);$$

(2) there exist two different values $a, b \in \mathcal{X}_1$, and a constant C, such that $\forall x \in \mathcal{X}_2$,

$$\Pr(X_1 = a \mid X_2 = x) = C \cdot \Pr(X_1 = b \mid X_2 = x).$$

Theorem 4.3. *Suppose that a categorical random variable $X_1 \in [T_1]$ and a discrete random variable $X_2 \in \mathbb{Z}$ are generated by a TSCM $((S_1, S_2), \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $((S_1, S_2), \mathcal{L}(X_1, X_2))$ satisfies Cond. 4.2, then $\mathcal{G}$ is identifiable.*

Proof. We need to prove that at most one of Eq. (A.16) and (A.17) hold:

$$X_1 = \sum_{j=1}^{K_1} \mathbb{I}_{\mathcal{R}_1^j}(X_2)\, f_1^j(N_1), \quad K_1 < |\mathcal{X}_2|, \quad \text{and} \quad X_2 \perp\!\!\!\perp N_1, \tag{A.16}$$

$$X_2 = \sum_{j=1}^{K_2} \mathbb{I}_{\mathcal{R}_2^j}(X_1)\, f_2^j(X_1) + N_2, \quad K_2 < T_1, \quad \text{and} \quad X_1 \perp\!\!\!\perp N_2. \tag{A.17}$$

Suppose that Eq. (A.16) holds. The following proof is based on Cai et al. [22]. Let $Z \triangleq h(X_2)$ be a random variable that represents the region X_2 falls into, meaning that $\mathbb{I}_{\mathcal{R}_1^Z}(X_2) = \mathbb{I}_{\mathcal{R}_1^{h(X_2)}}(X_2) = 1$. Since Z can take at most $K_1 < |\mathcal{X}_2|$ different values, there must exist two different a, $b \in \mathcal{X}_2$ such that $h(a) = h(b) = z_0$. We have

$$\frac{\Pr(X_1 = x_1)\Pr(X_2 = a \mid X_1 = x_1)}{\Pr(X_2 = a)} \tag{A.18}$$

$$= \Pr(X_1 = x_1 \mid X_2 = a) \tag{A.19}$$

$$= \Pr(X_1 = x_1 \mid X_2 = a, Z = z_0) \tag{A.20}$$

$$= \Pr(X_1 = x_1 \mid Z = z_0). \tag{A.21}$$

The last step is because of $X_1 \perp\!\!\!\perp X_2 \mid Z$. Similarly, we have

$$\frac{\Pr(X_1 = x_1)\Pr(X_2 = b \mid X_1 = x_1)}{\Pr(X_2 = b)} = \Pr(X_1 = x_1 \mid Z = z_0). \tag{A.22}$$

Combining the above two equations, we have

$$\Pr(X_2 = a \mid X_1 = x_1) = \Pr(X_2 = b \mid X_1 = x_1) \cdot \frac{\Pr(X_2 = a)}{\Pr(X_2 = b)}, \tag{A.23}$$

which satisfies (1) in Cond. 4.2 with $C = \frac{\Pr(X_2=a)}{\Pr(X_2=b)}$. Similarly, (2) in Cond. 4.2 is satisfied when Eq. (A.17) holds. Thus, we get a contradiction when both Eq. (A.16) and (A.17) hold, which concludes the theorem. □

Theorem 4.4. *Let $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_d)$ be generated from a TSCM with graph $\mathcal{G}$. If $\mathcal{L}(\mathbf{X})$ satisfies causal minimality [2], and for all $j \in \mathbf{V}$, $i \in \mathbf{PA}_j$ and all sets $\mathbf{S} \subseteq \mathbf{V}$ with $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i, j\}$, where $\mathbf{ND}_j$ represents the non-descendents of j, there is an $x_\mathbf{S}$ with $p_\mathbf{S}(x_\mathbf{S}) > 0$ such that $\left(S_{i|x_{\mathbf{PA}_j \setminus \{i\}}}, \mathcal{L}(X_i \mid X_\mathbf{S} = x_\mathbf{S}), \mathcal{L}(N_j)\right)$ satisfies the conditions for corresponding variable types in Cor. 4.1, Thm. 4.2-4.3, and Prop. B.1-B.3 (in Appendix, stating identifiability results for variables that share a common data type), then $\mathcal{G}$ is identifiable.*

Proof. The proof remains the same as that of Thm. 28 in Peters et al. [21]. As stated in Remark 30 in Peters et al. [21], the same proof can be directly applied to give valid multivariate identifiability whenever we have conditions that ensure bivariate identifiability since the proof is based on graphical causal structures instead of model assumptions. □

Proposition 5.1. *For any categorical variable X_i generated from Eq. (3.2) with non-empty parents, there exist a random variable $N_i' \not\perp\!\!\!\perp \mathbf{PA}_i$ and functions $\{g_i^j\}_{j=1}^{K_i}$ such that generating X_i using*

$$X_i := \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}(\mathbf{PA}_i)\, g_i^j(N_i')$$

gives identical observational distribution of X_i and $\mathbf{PA}_i$.

Proof. Let $X_j \in \mathbf{PA}_i$ be a parent of X_i. Let $N_i' = X_j + U$, where $U \sim \text{Gaussian}(0,1)$. It is immediate that $N_i' \not\perp\!\!\!\perp \mathbf{PA}_i$. We only need to show that there exists $\{g_i^j\}_{j=1}^{K_i}$ such that

$$\Pr(X_i' = x_i \mid \mathbf{PA}_i) = \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}(\mathbf{PA}_i) \Pr(g_i^j(N_i') = x_i) = \sum_{j=1}^{K_i} \mathbb{I}_{\mathcal{R}_i^j}(\mathbf{PA}_i) \Pr(f_i^j(N_i) = x_i) = \Pr(X_i = x_i \mid \mathbf{PA}_i). \tag{A.24}$$

The proposition follows by noting that the above requirement can always be met since we do not put restrictions on the form of $\{g_i^j\}_{j=1}^{K_i}$ and N_i' can take any value in $\mathbb{R}$. □

Proposition 5.2. *Suppose that $\mathcal{G}^*$ is the identifiable true causal graph generating $D{=}\{(x_1^k,\ldots,x_d^k)\}_{k=1}^m$, under some technical conditions [23, Thm. 1], the penalized log-likelihood score is consistent, i.e., $\forall\ \mathcal{G} \neq \mathcal{G}^*$,*

$$\Pr\left(s\left(\mathcal{G}, D\right) < s\left(\mathcal{G}^*, D\right)\right) \to 1 \quad \textit{as} \quad m \to \infty.$$

Proof. The technical conditions and proofs can be found in Nowzohour and Bühlmann [23, Thm. 1]. □

B Additional Identifiability Results

Here we give identifiability results for variables of a common data type, which are mainly based on previous works of Hoyer et al. [5], Peters et al. [7], Cai et al. [22].

Condition B.1. (Peters et al. [7]) *The tuple $((S_1, S_2), \mathcal{L}(X_1, X_2))$, where $X_1 \in \mathcal{X}_1 \subseteq \mathbb{Z}$ and $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, satisfies that either X_1 or X_2 has finite support. Let $X \in \{X_1, X_2\}$ denote the cause variable, $Y \in \{X_1, X_2\}$ denote the effect variable and f denote the assigned function in the causal mechanism. There does not exist a disjoint decomposition $\bigcup_{i=0}^{l} C_i = \operatorname{supp} X$ such that 1-3 are satisfied:*

1. *The C_i s are shifted versions of each other,*

$$\forall i\ \exists d_i \geq 0 : C_i = C_0 + d_i,$$

and f is piecewise constant: $f|_{C_i} \equiv c_i \forall i$.

2. *The probability distributions on the C_is are shifted and scaled versions of each other with the same shift constant as above: For $x \in C_i, \Pr(X = x)$ satisfies*

$$\Pr(X = x) = \Pr\left(X = x - d_i\right) \cdot \frac{\Pr\left(X \in C_i\right)}{\Pr\left(X \in C_0\right)}.$$

3. *The sets $c_i + \operatorname{supp} N := \{c_i + h : n(h) > 0\}$ are disjoint.*

Proposition B.1. (Peters et al. [7]) *Suppose that a discrete random variable $X_1 \in \mathbb{Z}$ and a discrete random variable $X_2 \in \mathbb{Z}$ are generated by a TSCM $((S_1,\ S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $((S_1, S_2), \mathcal{L}(X_1, X_2))$ satisfies Cond. B.1, then $\mathcal{G}$ is identifiable.*

Condition B.2. (Hoyer et al. [5]) *Given the tuple $((S_1, S_2), \mathcal{L}(X_1, X_2))$, where $X_1 \in \mathcal{X}_1 \subseteq \mathbb{Z}$ and $X_2 \in \mathcal{X}_2 \subseteq \mathbb{Z}$, let $X \in \{X_1, X_2\}$ denote the cause variable, $Y \in \{X_1, X_2\}$ denote the effect variable and f denote the assigned function in the causal mechanism. All related probability densities are strictly positive and that all densities, f_i and g_i are three times differentiable for $i = 1, 2, \cdots, k$. The following differential equation does not hold for any x and y with $\nu''(y{-}f(x))f'(x) \neq 0$:*

$$\xi''' = \xi''\left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}\right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' \left(f''\right)^2}{f'}, \tag{B.25}$$

where $\nu = \log p_{n_y}$, $\xi = \log p_x$, and we have skipped the arguments $y - f(x), x$, and x for ν, ξ, and f and their derivatives, respectively.

Proposition B.2. (Hoyer et al. [5]) *Suppose that a continuous random variable $X_1 \in \mathbb{R}$ and a continuous random variable $X_2 \in \mathbb{R}$ are generated by a TSCM $((S_1,\ S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $((S_1, S_2), \mathcal{L}(X_1, X_2))$ satisfies Cond. B.2, then $\mathcal{G}$ is identifiable.*

Condition B.3. (Cai et al. [22]) *The tuple $((S_1, S_2), \mathcal{L}(X_1, X_2))$, where $X_1 \in \mathcal{X}_1 = [T_1]$ and $X_2 \in [T_2]$, satisfies that if X_2 is the parent of X_1, the size of the parental partition $K_1 < T_2$ and that if X_1 is the parent of X_2, the size of the parental partition $K_2 < T_1$. Let $X \in \{X_1, X_2\}$ denote the cause variable, $Y \in \{X_1, X_2\}$ denote the effect variable. There does not exist two different values y_1, y_2, and a constant C, such that $\forall x \in \mathcal{X}$,*

$$\Pr\left(Y = y_1 \mid X = x\right) = C \cdot \Pr(Y = y_2 \mid X = x).$$

Proposition B.3. (Cai et al. [22]) *Suppose that a categorical random variable $X_1 \in [T_1]$ and a categorical random variable $X_2 \in [T_2]$ are generated by a TSCM $((S_1,\ S_2),\ \mathcal{L}(N_1, N_2))$ with a graph $\mathcal{G} : X_1 \to X_2$ or $\mathcal{G} : X_1 \leftarrow X_2$. If $((S_1, S_2), \mathcal{L}(X_1, X_2))$ satisfies Cond. B.3, then $\mathcal{G}$ is identifiable.*

C Experimental Details

C.1 Synthetic Data Generation To generate datasets from TSCMs, we specify the number of variables d, the number of variables of each type, and the number of edges e. In some experiments, we specify the graph density, defined as $\rho \triangleq 2e/d^2$, instead of directly specifying e. We randomly set the data type of each variable to be continuous, discrete, or categorical to satisfy the requirement. An undirected graph is sampled from the Erdős-Rényi model with the expected number of edges equal to e. We obtain a DAG by orienting the undirected edges following a uniformly randomly picked ordering of the d variables. Following the topological order in the DAG, we generate the value of each variable from the value of its parents and corresponding noise.

For variable X_i with a parental set $\mathbf{PA}_i$, we randomly split the sample space of each parent $X_j \in \mathbf{PA}_i$ into two disjoint subsets by randomly selecting a split point if X_j is a numerical variable and $|\mathcal{X}_j|$ disjoint subsets if X_j is categorical. Then all possible combinations of these subsets, form a parental partition for X_i. The generation process for variables of different types is described as follows:

1. If X_i is a continuous variable, the function f_i^j associated with each region in the partition is randomly selected from a linear function, or a nonlinear multilayer perceptron with sigmoid activation functions and two hidden layers. The parameters of f_i^j are uniformly chosen from $[-2, -0.5] \cup [0.5, 2]$. The noise variables are sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ where $\sigma \sim \text{Uniform}(0.1, 0.5)$.
2. If X_i is a discrete variable, the above procedure remains the same and we discretize it into b discrete values, where b is a uniformly sampled integer between 50 and 200.
3. If X_i is categorical, we implicitly specify the generating function. We randomly choose a category c_i^j for each region, then sample x_i from a categorical distribution that assigns c_i^j to x_i with probability p_c, where $p_c \sim \text{Uniform}(0.6, 0.9)$. If $\mathbf{PA}_i$ is empty, then the parental partition has only one region which is exactly the entire sample space and X_i is fully determined by the noise variables from the above procedures.

C.2 More Experimental Results We present the precision and recall score of each method in Fig. 5 and 6. The experimental settings remain the same as in Sec. 6.1. TRACER achieves significantly higher precision and recall score than other baselines. The GS method achieves higher recall than TRACER mainly because it outputs an equivalence class that contains undirected edges, and an edge is counted as correctly identified even if it is recognized as undirected.

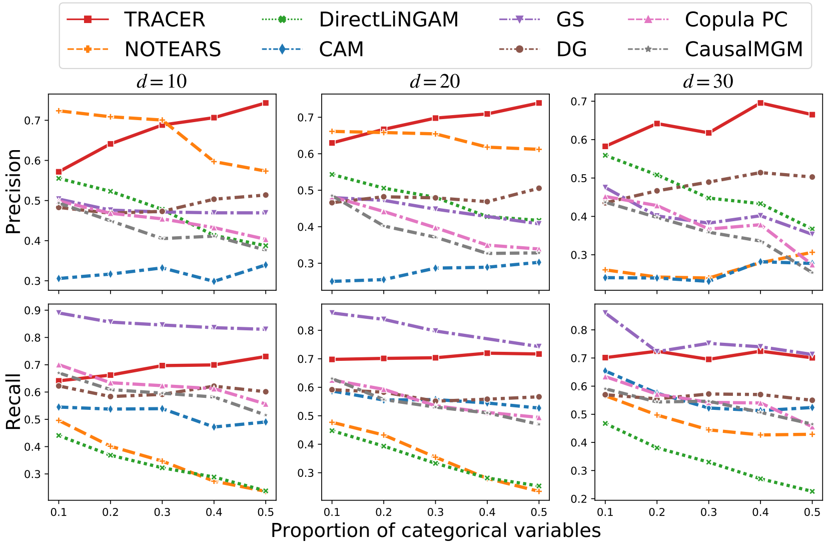


Figure 5: Averaged precision and recall over 100 simulations with varying proportions of categorical variables. The proportion of discrete variables is fixed at 0.3. For methods that output an equivalence class, the SID curve is plotted by averaging the corresponding upper and lower bound for better readability.

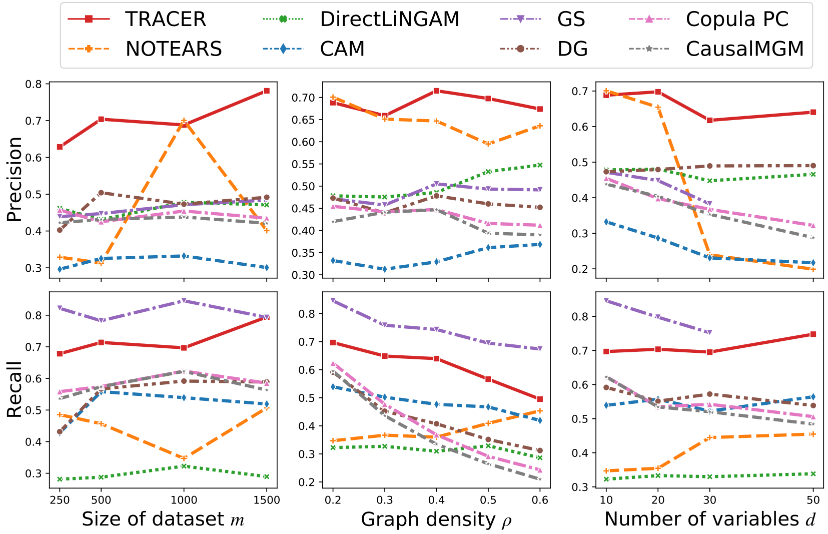


Figure 6: Averaged precision and recall over 100 simulations with varying $m/\rho/d$. The proportion of continuous/discrete/categorical variables are 0.4/0.3/0.3 respectively. Unless specified by the x-axis, we set $m = 1000$, $d = 10$, and $\rho = 0.2$.