Causal discovery in mixed additive noise models

Ruicong Yao[1,2] **Tim Verdonck**[2] **Jakob Raymaekers**[2]

[1]KU Leuven [2]University of Antwerp

Abstract

Uncovering causal relationships in datasets that include both categorical and continuous variables is a challenging problem. The overwhelming majority of existing methods restrict their application to dealing with a single type of variable. Our contribution is a structural causal model designed to handle mixed-type data through a general function class. We present a theoretical foundation that specifies the conditions under which the directed acyclic graph underlying the causal model can be identified from observed data. In addition, we propose Mixed-type data Extension for Regression and Independence Testing (MERIT), enabling the discovery of causal connections in real-world classification settings. Our empirical studies demonstrate that MERIT outperforms its state-of-the-art competitor in causal discovery on relatively low-dimensional data.

1 INTRODUCTION

Many real-world phenomena can be described by a set of variables and a graph whose directed edges encode the causal relationships between the variables. Causal discovery aims to identify causal graphs from observational or interventional data, which has many impactful applications in fields including medicine, economics, biology, and engineering. For example, pinpointing the chemical compounds in a drug that have curing effects or determining economic policies that drive inflation are crucial for informed decision-making. Beyond its direct applications, causal discovery is integral to domain generalization challenges, where a known causal structure guides variable selection to minimize generalization error [Christiansen et al., 2021].

Despite the many potentially impactful applications, identifying causality from observational data is highly non-trivial in practice. In fact, it can be shown that in full generality and without any restrictions, it is infeasible [Peters et al., 2017]. Therefore, many graph learning algorithms such as PC [Spirtes et al., 2000, Colombo et al., 2014] and GES [Chickering, 2002, Ramsey et al., 2017] focus on identifying the Markov equivalence class of the underlying graph, i.e. the set of all graphs which encode the conditional dependencies observed in the data. Conducting randomized experiments is one way to overcome this issue, but this is often prohibitive, unethical, or impossible. An alternative that has gained in popularity recently, is to add further restrictions to the causal models. By restricting the way the variables can be connected and/or generated, provable identifiability of the causal graph can be obtained [Shimizu et al., 2006, Peters et al., 2014, 2017]. A line of this research is based on the structural causal model (SCM) [Pearl, 2009]. In an SCM, each variable is determined by its causal parents and some mutually independent noise. There are restrictions on the function class and noise variables which prevent different SCMs within that class from generating the same joint distributions, thereby yielding identifiability.

SCMs with purely discrete or continuous variables have been studied quite intensively. For continuous variables, Shimizu et al. [2006] proved that causal models generated by linear functions and non-Gaussian noise (LiNGAM) are identifiable. Peters et al. [2014] showed that in general, identifiability holds for additive noise models. They also proposed the RESIT algorithm which is based on calculating the dependency between the predictors and the residuals. Immer et al. [2023] later extended the method to the heterogeneous noise causal model, Several score-based methods such as CAM [Bühlmann et al., 2014] and SCORE [Rolland et al., 2022] further explored the parametric structure of the additive noise model which enables efficient causal estimation. Zhang and Hyvärinen [2009] studied a more general functional model which includes an invertible link function outside of the additive noise structure, which also leads to structure identifiability. For discrete variables, Peters et al. [2011] proved the identifiability

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

for additive noise models between discrete variables. In particular, they assume that the variables and the noise are defined on a cyclic domain. Ni and Mallick [2022] showed that if the transition kernel between the discrete variables can be parametrized by ordinal regression, then the causal direction is also identifiable.

The study of causal discovery on mixed variables (both categorical and continuous) has received less attention, even though reality often confronts us with such data. For example, categorical variables can be the predictor in a regression problem and continuous variables can be the predictor in a classification problem. In addition, recent theoretical analysis on domain generalization is based on a causal graph between the causal/spurious features and the class labels [Arjovsky et al., 2019, Rosenfeld et al., 2021, Ahuja et al., 2021]. Therefore, addressing the causal modeling of such mixed-type datasets is necessary and would potentially be the foundation of complex causal learning tasks.

Related Work. A critical set of studies has begun to explore this area. Kocaoglu et al. [2017] considered the two-variable causal system, and derived identifiability guarantees for mixed-type data (via discretization) based on entropy constraints. Other research considered functional constraints such as a linear relationship between continuous response and each continuous variable. For example, Copula PC [Cui et al., 2016] assumes that the data is first generated from a linear Gaussian system, with categorical variables subsequently derived using a copula-based approach. The method then employs the Gaussian factor model to estimate the correlation matrix, and the PC algorithm to detect (un)directed edges of the graph. CausalMGM [Sedgewick et al., 2016] considers the linear Gaussian model (continuous) and the multinomial logistic regression model (categorical). It uses a likelihood-based method with ℓ_1 penalty to select causal parents. The conditional Gaussian score (CG) [Andrews et al., 2018] assumes the data comes from Gaussian mixtures to alleviate the score estimation in the method. Zeng et al. [2022] considered bi-level categorical variables modeled by a sign model applied to a linear transform and proposed a score-based algorithm LiM. In Liu et al. [2024], the authors considered mixed linear and nonlinear relations but focus on regression problems.

There are also methods that dropped the linearity assumption. For example, Andrews et al. [2018] proposed Mixed Variable Polynomial (MVP) score where the causal functions are modelled by polynomials. The max-min hill climbing method (MMHC) [Tsamardinos et al., 2006] on the other hand is a hybrid algorithm. In the first step, the PC algorithm is applied to learn the skeleton. Then, the directions are selected to maximize the likelihood greedily. However, as noticed in Li et al. [2022], normal conditional independence test might not perform well for mixed type data. In addition, for the second step, edges to categorical variables would always be preferred due to their smaller domain size. To overcome these issues, the authors proposed an algorithm HCM to solve the hybrid causal model. In particular, HCM extends the random conditional independence test [Strobl et al., 2019], a fast approximation of the kernel-based test [Zhang et al., 2011], to mixed-type data. In addition, they proposed a cross-validated BIC criterion to reduce the selection bias. Empirical studies showed that it outperformed MMHC for high-dimensional data. For the theory, they showed bivariate identifiability of the model and discussed the multivariate case. In [Huang et al., 2018] the authors proposed the Generalized score (GS) for mixed-type data based on a likelihood function for regression in RKHS and cross-validated selection. As a side remark, only LiM, HCM and GS in the above methods are designed to identify the causal graph under further assumptions. The others can only identify the Markov equivalence class.

Contributions. We consider the problem of causal discovery in structural causal models with both categorical and continuous variables, and provide general identifiability results. In addition, we generalize the approach of RESIT to detect the causal graph by comparing the residuals of model fits and the predictors. This is different from existing learning methods which are either score-based or constraint-based. Note that the former usually have parametric assumptions on the function class or the noise distribution, while the latter can only identify the Markov equivalent class.

Our contributions are:

- Propose a general function class and prove the identifiability of the structural causal models built on this class. In addition, we give specific conditions on the function class to be identifiable under certain noise distributions.

- Propose a new algorithm MERIT (Mixed-type data Extension for Regression and Independence Testing) for the classification setting which provably identifies the causal predictors.

- Illustrate the versatility of the model by generating different noise distributions.

- Demonstrate that MERIT outperforms the SOTA method on synthetic and real-world datasets in relatively low-dimensional scenarios.

2 BACKGROUND

To lay the foundation of our contributions, we will first define the core concepts and the general objective. Throughout, we assume that we have observations from a p-dimensional random vector $\boldsymbol{X} := \{X_1, \dots, X_p\}$ with distribution $\mathbb{P}_{\boldsymbol{X}} \subset \mathcal{P}_{\mathbb{R}^p}$. In addition, we assume that all related variables are observed, which is known as causal sufficiency.

Preliminaries. In order to consider structure identifiability, we need to assume a certain structure on the data-generating process underlying $\boldsymbol{X}$. In particular, we assume that the distribution $\mathbb{P}_{\boldsymbol{X}}$ is entailed by a structural causal model (SCM) $\mathcal{C}(\boldsymbol{S}, \mathbb{P}_{\boldsymbol{N}})$ [Peters et al., 2017]. The SCM is quantified by the pair $(\boldsymbol{S}, \mathbb{P}_{\boldsymbol{N}})$ where $\boldsymbol{S}$ is a set of structural assignments

$$X_i := f_i(\mathbf{PA}_i, N_i) \qquad i = 1, \dots, p.$$

Here $\mathbf{PA}_i \subset \{X_1, \dots, X_p\} \backslash \{X_i\}$ are the parents of X_i and $\mathbb{P}_{\boldsymbol{N}}$ denotes the joint distribution over the noise variables $N_1, \dots, N_p$ which are assumed to be jointly independent. The graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ of the SCM is a directed acyclic graph (DAG) with nodes $\mathcal{V} = \{1, \dots, p\}$ and directed edges $\mathcal{E} = \{i \to j, \text{if } i \text{ causes } j,\ i,\ j \in \mathcal{V}\}$, which can be obtained by drawing one node for each X_i and drawing edges from $\mathbf{PA}_i$ to X_i. We use $\mathcal{F}$ to denote the class of functions f_i under consideration. In particular, we use $\mathcal{F}|_2$ to denote the function class on $(X_j, N_i), \forall j$ formed by $f_i(x_{\mathbf{PA}_i \backslash \{j\}}, \cdot, \cdot)$.

Identifiable Causal Discovery. Causal discovery aims to recover the graph $\mathcal{G}$ underlying data-generating process based on i.i.d. samples from the distribution $\mathbb{P}_{\boldsymbol{X}}$. It is well-known that this is only possible under further assumptions [Peters et al., 2017]. One way to obtain identifiability is through an a priori restriction of the model class. The most common way of achieving this is by limiting the complexity of the functions in $\mathcal{F}$, and we take this approach as well. We first consider bivariate identifiability, i.e. can we infer from the data of two random variables X, Y whether $X \to Y$ or $Y \to X$ where the connection of the two variables is given by a function in $\mathcal{F}$. In case this is possible, the functions in $\mathcal{F}$ form a *bivariate identifiable set* $\mathcal{B} \subset \mathcal{F}|_2 \times \mathcal{P}_{\mathbb{R}} \times \mathcal{P}_{\mathbb{R}}$. Formally, we can define the Bivariate Identifiable Set [Peters et al., 2012] as Definition 3. Bivariate identifiability can be leveraged in the identifiability of the underlying graph in the multivariate case as Theorem 4, under the assumption that the SCM with function class $\mathcal{F}$ is an Identifiable functional model class ($(\mathcal{B}, \mathcal{F})$-IFMOC, Definition 4). We refer to Appendix B.1 for more details.

3 MIXED ADDITIVE NOISE MODELS

In this section, we define models for the continuous and categorical variables. For the former, we use the additive noise model. More precisely, if X_i is a continuous variable we assume

$$X_i := f_i(\mathbf{PA}_i) + N_i. \tag{1}$$

Here, f_i can be both linear or nonlinear, and N_i is assumed to have density p_{N_i}. To be clear, such linearity is with respect to the continuous predictors. For each categorical predictor with K levels, f_i is simply a map from $\{1, \dots, K\}$ to $\{f_i(1), \dots, f_i(K)\}$ given other predictors of X_i.

For modeling categorical variables, we use the K-level argmax model, i.e., for $\forall X_i$ categorical, we assume

$$X_i := \underset{k \in \{1, \dots, K\}}{\arg\max} \left(f_{i,k}(\mathbf{PA}_i) + N_{i,k}\right) \tag{2}$$

with $N_{i,k}$ mutually independent. For the $f_{i,k}$, we use the same classes of functions as for the continuous variables. The noise term inside the $\arg\max$ yields a flexible model where the noise can differ between categories. We believe this is more practical than assuming homogeneous noise as the signal-noise ratio is usually smaller around the decision boundary. It is interesting to note that a special case of this model, where $N_{i,k} \sim \text{Gumbel}(0, 1)$ for all $k = 1, \dots, K$, can be equivalently written as the multinomial logistic model (see Appendix D.4 for a proof)

$$P(X_i = k | \mathbf{PA}_i) := \frac{\exp(f_{i,k}(\mathbf{PA}_i))}{\sum_k \exp(f_{i,k}(\mathbf{PA}_i))}.$$

This specific model was also used by Li et al. [2022]. For other noise distributions, we do not have explicit expressions for the probability distribution in general. Another special case of this model is the choice $K = 2$ (i.e. a binary categorical variable) with $N_{i,1}, N_{i,2}$ chosen such that $N_{i,1} - N_{i,2} \sim \tilde{N}_i$. In that case, we recover the model $X_i := \text{sign}(f_i(\mathbf{PA}_i) + \tilde{N}_i)$ suggested by Zeng et al. [2022]. It should be clear that the proposed model includes many popular models for categorical variables but is substantially more general.

As a side remark, in causal discovery we are primarily interested in the identification of the causal graph. However, in a general causal inference context we may additionally be interested in identifying the precise functions f_i in the argmax model. Without any further restrictions, given the knowledge of the conditional distribution, they can only be identified up to a constant. To make f_i unique, we would need a mild additional restriction. More precisely, we would either assume that f_1 is fixed to a constant or $\sum_i f_i(X) \equiv 0$ which is common in the literature. In this case, the model can also answer some counterfactual queries, see our discussions in Appendix C. Overall, we define the entire causal model as follows.

Definition 1 (Mixed additive noise model)**.** *A SCM with function class $\mathcal{F} = \mathcal{F}^c \cup \mathcal{F}^d$ is called a mixed additive noise model (MANM) if $\mathcal{F}^c$ and $\mathcal{F}^d$ induce (see Equation (8) in Appendix B.1) the respective bivariate function classes $\mathcal{F}|_2^c$ and $\mathcal{F}|_2^d$ where*

$$\begin{aligned} \mathcal{F}|_2^c &= \{f(X) + N\}\,, \\ \mathcal{F}|_2^d &= \left\{ \underset{k\in\{1,\ldots,K\}}{\arg\max}\ (f_k(X) + N_k) \right\}. \end{aligned} \tag{3}$$

Here $N, (N_k)_{k\in\{1,\ldots,K\}}$ are real random variables, and X can be either continuous or categorical. The superscript c and d indicate whether the response is continuous or discrete (categorical).

3.1 Multivariate Identifiability

Suppose the variables $\{X_1, \ldots, X_p\}$ satisfy our SCM of Definition 1, we want to show that the causal graph of a MANM is identifiable from i.i.d. samples of the joint distribution of the variables. We start by considering bivariate identifiability, i.e. the case of two random variables X and Y. We distinguish between three cases. The first is when both X and Y are continuous. In that case, bivariate identifiability follows from the identifiability of additive noise models established in Shimizu et al. [2006], Peters et al. [2014]. For the second case where we have categorical pairs, the following holds

Theorem 1. *Let X, X', Y, Y' be categorical variables. Let $\mathbb{P}_{(X,Y)}$ and $\mathbb{P}_{(X',Y')}$ be induced by Equation (2):*

$$\begin{aligned} Y &:= \underset{1\le k_y\le K_y}{\arg\max}(f_{k_y}(X) + N_{k_y}), \\ X' &:= \underset{1\le k_{x'}\le K_{x'}}{\arg\max}\ (f'_{k_{x'}}(Y') + N'_{k_{x'}}), \end{aligned}$$

such that $f_{k_1}, f'_{k_1} \equiv 0$ and f_{k_y}, $f'_{k_{x'}}$ not all constant. N_{k_y} and $N'_{k_{x'}}$ have density $p_{N_{k_y}}$ and $p_{N'_{k_{x'}}}$ with distribution function $F_{N_{k_y}}$ and $F_{N_{k_{x'}}}$ and are mutually independent. Then, if $\mathbb{P}_{(X,Y)} = \mathbb{P}_{(X',Y')}$, we must have for all (k_x, k_y, i_x, i_y) that

$$\begin{aligned} &\frac{P(X = k_x|Y = k_y)P(X = i_x|Y = i_y)}{P(X = k_x|Y = i_y)P(X = i_x|Y = k_y)} \\ &= \frac{P'(X' = k_x|Y' = k_y)P'(X' = i_x|Y' = i_y)}{P'(X' = k_x|Y' = i_y)P'(X' = i_x|Y' = k_y)}, \end{aligned} \tag{4}$$

where the conditional distribution is induced by

$$P(Y = k_y|X = k_x) = \int_{\mathbb{R}} p_{N_{k_y}}(z) \prod_{k'_y\neq k_y} F_{k_y}(f_{k_y}(k_x) - f_{k'_y}(k_x) + z)dz.$$

Therefore, the causal direction is identifiable unless condition (4) holds.

Special cases were established in Zeng et al. [2022], Li et al. [2022], but with the restriction that all noise variables follow the Gumbel distribution. In the following, we provide a concrete example where condition (4) is induced by functional and distributional constraints. Consequently, our results also incorporate the multinomial probit model (where N_{k_x} is Gaussian), which is popular in the econometric literature [McCulloch et al., 2000]. All the proofs will be deferred to Appendix D.

Corollary 1. *Consider a model between two binary variables $X \to Y$ given by*

$$Y = \underset{\alpha\in\{0,1\}}{\arg\max}((1-\alpha)N_{k_0} + \alpha(c + \beta_x X + N_{k_1})),$$

where N_{k_0}, N_{k_1} are the continuous noise, $c \in \mathbb{R}$ is a known constant. Then, the model is identifiable if X and Y do not have the same marginal distribution. That is, $\nexists(X', Y')$, s.t. $\mathbb{P}_{(X,Y)} = \mathbb{P}_{(X',Y')}$,

$$X' = \underset{\alpha\in\{0,1\}}{\arg\max}((1-\alpha)N'_{k_0} + \alpha(c + \beta_y Y' + N'_{k_1})).$$

The third case, the model of a categorical and a continuous variable, is the most interesting one. Theorem 2 confirms the bivariate identifiability of the continuous-categorical case given a mild condition.

Theorem 2. *Let X, X' be continuous variables and Y, Y' be categorical variables. Let $\mathbb{P}_{(X,Y)}$ be induced by Equation (2), that is*

$$Y := \underset{k_y\in\{1,\ldots,K_y\}}{\arg\max}\ (f_{k_y}(X) + N_{k_y}),$$

where K_y is the number of levels, and N_{k_y} are i.i.d. random variables independent of X with distribution function F_{k_y} and density $p_{N_{k_y}}$. We assume that $f_{k_1} \equiv 0$ and at least one f_{k_y} is not a constant. Consider any distribution $\mathbb{P}'_{(X',Y')}$ induced by the reverse SCM s.t.

$$X' := f'_{X'}(Y') + N_{X'}$$

where $N_{X'} \perp\!\!\!\perp Y'$, $N_{X'} \sim \mathbb{P}_{N_{X'}}$ with density $p_{N_{X'}}$ and $f'_{X'}$ not a constant. Then, if $\mathbb{P}_{(X,Y)} = \mathbb{P}_{(X',Y')}$, we must have that for $\forall 1 \le k_y \le K_y$, $\forall x :$

$$P(Y = k_y|X = x) = P'(Y' = k_y|X' = x), \tag{5}$$

where

$$P(Y = k_y|X = x) = \int_{\mathbb{R}} p_{N_{k_y}}(z) \prod_{k'_y \neq k_y} F_{k_y}(f_{k_y}(x) - f_{k'_y}(x) + z)dz,$$

$$P'(Y' = k_y|X' = x) = \frac{p_{k_y} p_{N_{X'}}(x - c_{k_y})}{\sum_{1 \le k'_y \le K_y} p_{k'_y} p_{N_{X'}}(x - c_{k'_y})}.$$

Here, $p_{k_y} = P(Y = k_y)$, and $c_{k_y} = f'_{X'}(Y' = k_y)$. Hence, the causal direction is identifiable unless condition (5) holds.

Remark 1. *Equation (5) is a very specific relation between the shifted distribution $p_{N_{X'}}(x - c_{k_y})$ and $f_{k_y}(x)$. Thus, we can identify the causation in two directions: 1. $X \to Y$ if $f_{k_y} \in \mathcal{F}$ can be estimated for some function class $\mathcal{F}$ and (5) can not hold under the assumption of N'_X. This is natural as the noise distribution is unlikely to follow a specific formula. 2. $Y \to X$ if f_{k_y} solving (5) does not exist or does not belong to the presumed function class $\mathcal{F}$. We note that when the truth is $Y \to X$, a backward model $X \to Y$ may exist for some f_{k_y} without restrictions on $\mathcal{F}$, e.g., when X is a mixture of Gaussians, see Corollary 2. Thus, to ensure identifiability, $\mathcal{F}$ must exclude such functions. In fact, such treatment on the function class (and thus the preference on the causal direction) is aligned with previous study [Janzing and Schölkopf, 2010], where $Y \to X$ was also preferred for its explanation on the multi-modal distribution and low Kolmogorov complexity.*

For certain noise distributions, we have even simpler criteria and can obtain the exact class of functions for which the the direction $Y \to X$ should be preferred.

Corollary 2. *If N_{k_y}, $1 \le k_y \le K_y$ follows a Gumbel(0,1) distribution, then Equation (5) is equivalent to require the following holds up to a constant*

$$f_{k_y}(x) = \log\left(\frac{p_{k_y}}{p_1}\right) + \log\left(\frac{p_{N_{X'}}(x - c_{k_y})}{p_{N_{X'}}(x - c_{k_1})}\right). \qquad (6)$$

If we further assume that $N_{X'}$ is Gaussian, then the model is identifiable if at least one f_{k_y} is non-linear.

A similar result was shown by Li et al. [2022] under the Gumbel noise assumption and a relatively complicated criterion with an integral over the discrete variables. Instead, Equation (6) presents a simple criterion to characterize the relationship between the functions and additive noise variables for the non-identifiable case.

When the domain of the variable X is $\mathbb{R}$, Equations (5) can be simplified using the asymptotic behavior of f_{k_y} and $N_{X'}$, as stated in Corollary 3. In fact, the theorem covers many interesting cases. For example, the first condition holds for Gaussian distributions. The second condition covers distributions like the Laplace distribution or the Gamma distribution.

Corollary 3. *Let X be a real random variable. Assume $p_{N_{X'}}$ and $p_{N_{k_y}}$ have a positive density on $\mathbb{R}$. The causal direction is identifiable if one of the following holds:*

1. *$\forall f_{k_y}, f_{k'_y}$: $\sup_x |f_{k_y}(x) - f_{k'_y}(x)| < \infty$, and $\forall a, b$ with $a < b$: $\lim_{x\to\infty} \frac{P_{N_{X'}}(x-a)}{P_{N_{X'}}(x-b)} = 0$.*
2. *$\exists f_{k_y}, f_{k'_y}$: $\lim_{x\to\infty} |f_{k_y}(x) - f_{k'_y}(x)| = \infty$, and $\forall a, b$ with $a < b$: $\lim_{x\to\infty} \frac{P_{N_{X'}}(x-a)}{P_{N_{X'}}(x-b)} = C \neq 0$ or the limit does not exist.*

We now obtain the bivariate identifiability of the function class stated in the following proposition

Proposition 1. *(The model set of MANM forms a bivariate identifiable set) The following sets $\mathcal{B}_1$ and $\mathcal{B}_2$ are bivariate identifiable sets follow from MANM:*

$$\begin{aligned} \mathcal{B}_1 :=& \{f \text{ linear and } N \text{ non-Gaussian in } \mathcal{F}|_2^c\} \cup \\ & \{f \in \mathcal{F}|_2^d \text{ with Condition } (4), (5) \text{ not hold}\} \\ \mathcal{B}_2 :=& \{f \text{ nonlinear in } \mathcal{F}|_2^c\} \cup \\ & \{f \in \mathcal{F}|_2^d \text{ with Condition } (4), (5) \text{ not hold}\} \end{aligned}$$

Proof. This immediately follows from Theorem 1,2, and the identifiability results of linear and nonlinear additive noise models. □

Now we turn to multivariate identifiability. The key observation is that here the bivariate identifiable set does not depend on the distribution of X itself which is different from the result in Li et al. [2022]. Thus, the restricted case in Equation (9) would not be affected by S. We then obtain the following theorem for the identifiability in the multivariate case.

Corollary 4. *(Multivariate identifiability) Let $f|_2$ be the bivariate restriction of f on variable X and noise N. If the distribution of $\{X_1, \ldots, X_p\}$ is generated by a $(\mathcal{B}, \mathcal{F})$-IFMOC, where $\mathcal{B} = \mathcal{B}_1$*

$$\begin{aligned} &\mathcal{F} = \{f \in \mathcal{F}^c \text{linear on predictors, } N \text{ non-Gaussian}\} \\ &\{f \in \mathcal{F}^d \text{ s.t. Condition } (4), (5) \text{ not hold.}\} \text{ or} \\ &\mathcal{F} = \{f \in \mathcal{F}^c \text{ nonlinear with additive noise}\} \cup \\ &\{f \in \mathcal{F}^d \text{ s.t. } f_k \text{ nonlinear with } (4), (5) \text{not hold}\}, \end{aligned}$$

with $\mathcal{B} = \mathcal{B}_2$, then the causal DAG is identifiable.

4 LEARNING ALGORITHM

In this section, we introduce our learning algorithm. Basically, causal discovery algorithms fall into the following categories: constraint-based (PC), score-based (CAM, SCORE), independent component analysis (ICA)-based and regression-based (RESIT). Existing works mainly focus on constraint-based or score-based approaches. ICA is primarily used for LiNGAM, although extensions to nonlinear causal models have also been proposed recently [Monti et al., 2020]. Regression-based methods have been explored less. However, since many real-world problems can be described by additive noise models and this is essentially the most flexible model with identifiability guarantees, it is still worthwhile to leverage these model structures in causal discovery. Moreover, the regression-based method only requires causal minimality to be identifiable instead of the faithfulness condition which is hard to verify [Peters et al., 2014]. Although score-based methods can also deal with additive noise models, they often require specific assumptions on the function class and the noise type [Bühlmann et al., 2014, Nowzohour and Bühlmann, 2016]. In contrast, RESIT has no such requirements. Therefore, we want to generalize RESIT which iteratively learns the causal structure by computing the dependency between the residuals and the predictor. For clarity, we will denote the continuous and categorical variables by X_i, Y_i respectively.

Mixed-type data Extension for Regression and Independence Testing. Categorical variables cannot be incorporated directly into the RESIT framework as there are no residual estimates of N_{k_y} in the classification. As a remedy, we propose MERIT, a 3-phase method as an extension of RESIT to mixed-type data. In its first phase, Algorithm 1 learns the causal order of the continuous variables and the relative order between the continuous variables and categorical variable blocks under the MANM. Suppose we have p_1 continuous variables $X_1, \ldots, X_{p_1}$ and p_2 categorical variables $Y_{p_1+1}, \ldots, Y_{p_1+p_2}$, where $p_1 + p_2 = p$. In order to construct our algorithm, we introduce a new concept to deal with the potential presence of contiguous categorical variables in the true causal order.

Definition 2 (Categorical block). *Let $B_Y \subset \{p_1 + 1, \ldots, p_1 + p_2\}$. B_Y is a categorical block if its elements appear contiguously in the true causal order, i.e. $\exists i, j \in \{1, \ldots, p_1\}$ with $i \neq j$ (i or j can be empty) so that*

$$\pi_{true} = [\ldots, i, B_Y, j, \ldots],$$

where π_{true} denotes the true causal order. We also use B_{Y_i} to denote the block that Y_i belongs to.

We give the general idea here, before showing the theoretical support. If a continuous variable X_i is the leaf node, then regressing X_i on all the remaining variables would always give independent residuals. This is because $X_i \perp\!\!\!\perp X_k | \mathbf{PA}(X_i)$ for any $X_k \notin \mathbf{PA}(X_i)$. In practice, independence is verified by observing a p-value larger than a given threshold, e.g. 0.05. Note that if a categorical variable Y_i and X_i are both the leaf nodes in the graph, X_i will be first considered and removed by design. Therefore, Algorithm 1 runs in the same way as RESIT until there are only categorical leaf nodes in the subgraph. This corresponds to the case that all the residuals tested in lines 6-7 are dependent when $S_Y = Re_Y$. We then want to find the proper categorical leaf nodes to drop so that some X_i would become a leaf node. This is done by iterating on the subsets of Re_Y, until we find the correct independent residuals (line 10). Then we recognize $B_Y = Re_Y \setminus S_Y$ as the indices of the variables to drop, whose causal order is below X_i, and repeat the process.

Algorithm 1 Causal order discovery

```
1: Input: I.i.d. samples from a $p$-dimensional distribution $(X_1, \ldots, X_{p_1}, Y_{p_1+1}, \ldots, Y_{p_1+p_2})$, $X_i$ continuous, $Y_i$ categorical, Test threshold $\alpha$.
2: $S := \{1, \ldots, p_1\}$ only contains the label of continuous variables, $\pi := [\ ]$ causal order of $X$ and $Y$ blocks ($B_Y$), $Re_Y = \{p_1 + 1, \ldots, p_1 + p_2\}$ labels of the remaining categorical variables.
3: repeat
4:   for $S_Y \in$ [subsets of $Re_Y$ (size decreasing)] do
5:     for $k \in S$ do
6:       Regress $X_k$ on $\{X_i\}_{i \in S \setminus \{k\}} \cup \{Y_i\}_{i \in S_Y}$.
7:       Measure dependence between residuals $r_k$ and $\{X_i\}_{i \in S \setminus \{k\}} \cup \{Y_i\}_{i \in S_Y}$.
8:     end for
9:     Let $k^*$ be the index of the variable with the weakest dependence.
10:    if the p-value regarding $X_{k^*}$ is larger than $\alpha$ then
11:      $B_Y = Re_Y \setminus S_Y; Re_Y = S_Y$
12:      $\mathbf{PA}(k \in B_Y) = S$ (continuous PA of $k$)
13:      $\mathbf{PA}(k^*) = S \setminus \{k^*\} \cup S_Y$; $\pi = [k^*, B_Y, \pi]$
14:      break
15:    end if
16:  end for
17:  $S = S \setminus \{k^*\}$ or $S = \{\}$ if $S = \{k_{\text{last}}\}$.
18: until $\#S = 0$
19: if $\{1, \ldots, p_1\} \subset \pi$ then
20:   $\pi = [Re_Y, \pi]$; $\mathbf{PA}(k \in Re_Y) = \{\}$
21: else
22:   $\pi = [k_{\text{last}}, Re_Y, \pi]$; $\mathbf{PA}(k_{\text{last}}) = \{\}$
23: end if
24: Output: $\pi, \{\mathbf{PA}(k), 1 \leq k \leq p\}$.
```

Once the topological ordering is found, we identify the parents of the continuous and the categorical variables in phases 2 and 3, respectively. The second phase of MERIT is identical to that of the RESIT [Peters et al., 2014] which prunes parents of continuous variables by conducting independent tests on the residuals and the predictors. This also results in $\mathbf{CH}_c(k)$, the continuous children of all variables. For the third phase, we start by using conditional independence tests to find the parents of Y_i that are continuous variables. The idea is that if $X_i \notin \mathbf{PA}(Y_j)$, then there must exist a set S whose causal order is equal or higher than B_{Y_j} s.t. $X_i \perp\!\!\!\perp Y_j | S$. Then, it remains to identify the causal relationship within the Ys, for which we can use the PC or the BDeu score [Buntine, 1991] to find the skeleton and then apply conditions on the function class to decide the direction, see Algorithm 2. We will show in Section 5 that such constraints are necessary for identifying causal directions. We also illustrate Algorithms 1 and 2 on a concrete example in Appendix E.

Theoretical Guarantees. Our algorithm is guaranteed to identify the true causal graph under certain assumptions. The first assumption is the causal minimality, which is generally weaker than faithfulness. The second assumption specifies requirements on the noise distribution N_X of the continuous variables and includes a broad class of distributions [Yakowitz and Spragins, 1968]. This identifiability requirement can be relaxed, which we discuss in Remark 2 in Appendix D.6. The third assumption is a variant of Assumptions 1, 2 in Maeda and Shimizu [2021]. In particular, its second condition is natural as Y_i satisfies the argmax model.

Assumption 1. *We assume that the observed distribution is Markov to the causal graph $\mathcal{G}$ and not Markov to any subgraph $\mathcal{G}'$ of $\mathcal{G}$.*

Assumption 2. *We assume that the class of the noise distributions N_X for the continuous variables satisfies: 1. Finite mixture of N_X is identifiable. 2. N_X itself is not a mixture of distributions in the class.*

Assumption 3. *We assume 1. f_X is additive on each categorical parent. 2. The residuals of $\forall X_i, Y_i$ regressed on any variables are dependent with N_{X_i}, N_{k_y}. 3. Let $G_i(M_i)$ be the regression function of X_i (continuous) on some set M_i. Then $\exists N = N_{X_i}$ (or (N_{k_y})) s.t. $N \not\perp\!\!\!\perp X_i - G_i(M_i) \wedge N \not\perp\!\!\!\perp X_j - G_j(M_j) \Rightarrow X_i - G_i(M_i) \not\perp\!\!\!\perp X_j - G_j(M_j)$. 4. $N_{X_i} \not\perp\!\!\!\perp Y_j$ if Y_j is X_i's descendant.*

Theorem 3. *Suppose the causal model satisfies MANM and the assumptions hold. Then on a population level, Algorithm 1 finds the correct causal order of X_i, $1 \leq i \leq p_1$, and all the categorical blocks B_Y. Algorithm 2 finds the correct continuous parents of the categorical variables. Moreover, if faithfulness is further assumed for line 8 (e.g. the PC algorithm), then it finds the correct parents of the categorical variables.*

The proof consists of several steps and can be found in Appendix D.6. One of the main tools is Theorem 34 of Peters et al. [2014]. The key observation is that our identifiability result in Theorem 2 only depends on the noise distribution and the function class, rather than the variable distribution. Thus restricted bivariate identifiability holds. Assumption 2 helps to deal with categorical confounders and Assumption 3 eliminates undesired results (line 10) due to the choice of S_Y. The justification of Algorithm 2 follows from the correct topological ordering and knowledge of $\mathbf{CH}_c(k)$.

Algorithm 2 Pruning for the categorical variable

```
 1: Input: I.i.d. samples from a $p$-dimensional distribution $(X_1, \ldots, X_{p_1}, Y_{p_1+p_2}, \ldots, Y_{p_1+p_2})$, $\pi$, threshold $\alpha$, $\mathbf{PA}(k), \mathbf{CH}_c(k), 1 \leq k \leq p$ after phase 2.
 2: for $p_1 < i \leq p_1 + p_2$ causal order bottom up do
 3:     for $k \in \mathbf{PA}(i), 1 \leq k \leq p_1$ do
 4:         Test the conditional independence between $Y_i$ and $X_k$ given $(CH_c(k) \backslash CH_c(i)) \cup S_{-Y_i}$ where $S_{-Y_i} \subset B_{Y_i} \backslash \{Y_i\}$ s.t. $Y_i \in B_{Y_i}$.
 5:         If exists p-value $> \alpha$, remove $k$ from $\mathbf{PA}(i)$.
 6:     end for
 7: end for
 8: Apply PC/BDeu for full connections between $Y_i$.
 9: Assert $Y_i \to Y_j$ if the categorical block $B_{Y_i}$ of $Y_i$ has higher order than the block $B_{Y_j}$ of $Y_j$ in $\pi$.
10: Use functional constraints to learn causal order of variables in each block $B_Y$.
11: Output: Full PA of each $Y_i$.
```

5 EXPERIMENTAL RESULTS

In this section, we evaluate our approach against other methods in the field of causal discovery involving mixed data types. Interestingly, HCM and GS stand out as the sole method for multi-level categorical variables as discussed in Li et al. [2022], Huang et al. [2018], whereas LiM can identify causal directions for bi-level categorical variables. Although there are alternative strategies for handling mixed-type data, such as causalMGM, Conditional Gaussian Score, and MMHC, these methods are limited to identifying the Markov equivalence class (MAC), which fall short of causal learning objectives. Since HCM and GS have demonstrated superior performance in various experiments, we focus on comparing MERIT with them. We also conducted experiments to test the necessity of a restricted functional class for identifying causal directions on bi-level categorical data, where PC and LiM are also included.

Parameters Setting. For MERIT, we use a significance level of $\alpha = 0.05$ for obtaining the topological ordering (i.e., Algorithm 1) and we use $\alpha = 0.01$ for the pruning procedure (phase 2 of MERIT and Algorithm 2). For the joint independence test between the variables and the residuals, we use a kernel-based test dHSIC [Pfister et al., 2018]. For the conditional independence test used in Algorithm 2, we use the cdcov.test from Wang et al. [2015] with parameter

$\alpha = 0.01$. Typically, α is a tuning parameter [Colombo et al., 2014]. When the underlying data is nonlinear, we use random forest regressor [Liaw and Wiener, 2002]. For HCM[1] we use default parameters for the tests. For GS, we used the implementation in Zheng et al. [2024]. For LiM, we use the implementation in Ikeuchi et al. [2023] with recommended parameters. For PC, we use pc.stable from `R` package `bnlearn` [Scutari, 2010] with mutual information test and $\alpha = 0.05$.

Synthetic Categorical Datasets We provide simulations on datasets with only categorical data to illustrate lines 8-10 of Algorithm 2 and justify the necessity of functional constraints. We considered dimension $p = 3, 5, 10$, number of samples $n = 1000$, and simulated Erdös-Renyi (ER) graphs [Frieze and Karoński, 2016] with probability 2/3 for each directed edge to existing and repeated 100 times. The data was simulated according to Equation (2) with $K = 2$, $N_{i,k} \sim \text{Gumbel}(0, 1)$ and f linear with intercept 0. Thus by Corollaries 1 and 4, the graphs are identifiable given the prior knowledge of the function and distribution. MERIT starts by applying the PC algorithm to learn the skeleton G_0. Then, it leverages the prior knowledge to identify the causal directions in the following steps. 1. Fit a logistic regression model for each variable with neighbors (other variables with undirected edges to it) in G_0. 2. Identify the variable whose intercept is closest to zero as the leaf node. 3. Include all its neighbors as its parents. 4. Remove it and repeat 2, 3 until all edges are directed. The results of MERIT, HCM, PC, LiM, and GS are shown in Table 1. METRIC clearly outperforms all methods. Moreover, other score-based methods do not substantially outperform PC, which only learns the MAC. This indicates that known functional constraints are necessary to ensure the identifiability of DAGs on categorical variables.

Table 1: Results on categorical datasets.

Metric	Method	$p = 3$	$p = 5$	$p = 10$
SHD	**MERIT**	**0.23±0.53**	**1.72±1.56**	**18.29±4.52**
	HCM	1.16±0.8	3.91±1.71	20.0±3.99
	PC	1.58±1.01	4.45±1.87	21.76±4.47
	LiM	1.46±0.82	5.08±1.44	25.14±3.66
	GS	1.9±0.89	4.82±1.79	20.57±4.71
SID	**MERIT**	**0.39±1.10**	**5.06±5.23**	**65.68±11.55**
	HCM	2.45±2.10	12.76±4.88	74.31±6.89
	PC	3.59±2.31	14.26±4.65	76.01±8.22
	LiM	2.63±2.24	13.43±4.20	76.49±6.94
	GS	4.15±2.25	14.53±4.92	72.73±8.52
F1	**MERIT**	**0.85±0.32**	**0.80±0.20**	**0.48±0.12**
	HCM	0.45±0.38	0.44±0.23	0.41±0.11
	PC	0.56±0.27	0.49±0.22	0.39±0.12
	LiM	0.37±0.35	0.37±0.19	0.27±0.09
	GS	0.54±0.24	0.54±0.20	0.45±0.13

Synthetic Datasets with Mixed Data Types. Here we consider two classes of graphs in the simulation: The ER graph where each position in the ordered (upper triangular) adjacency matrix has probability $2/(p-1)$ (p for the dimension) to be 1, and the Scale-Free graphs with parameter $m = 2$. Each dataset has 1000 samples and $p = 3$, 5, 7, or 10 (respectively 1, 2, 3, 5 categorical variables with 5 levels). We choose these dimensions to be compatible with the original RESIT which shows good performance for relatively low-dimensional data. We do not consider the connection between the categorical variables because Table 1 has shown that even for simple DAGs on bi-level categorical variables, existing methods do not perform well as they do not consider functional constraints. Thus we do not want this drawback to affect the comparison in identifying the relationships between the continuous and categorical variables. We also provide further discussions in Remark 3 in Appendix E.

We generated both linear and nonlinear SCMs with identifiable causal structures. For linear functions, the coefficients were randomly sampled from $[-2, -0.5] \cup [0.5, 2]$. For the noise distribution in Equation (1), we either used $\text{Uniform}(-3, 3)$ or $\text{Logistic}(0, 2)$. For that in Equation (2), we used the Gumbel distribution. The categorical variable in this case changes the intercept of the function to continuous variables. For nonlinear functions, we simulated 2-layer Multi-layer perceptrons (MLPs) with hidden dimensions equal to 500. The choice of the noise distribution was the same as above. When a categorical variable is the input of the MLP, it is one-hot encoded. This also allows us to test the robustness of our algorithm when Assumption 3 is potentially violated. Note that here for simplicity, we let the function class in Equation (1) and (2) be the same. For each choice of p, class of functions, and noise distribution, we generated 20 datasets and calculated the average structural Hamming distance (SHD), structural intervention distance (SID) [Peters and Bühlmann, 2015] and F1 score of all the methods. Note that by convention, getting the direction wrong counts as a 1 in the SHD.

We first consider the results on the Erdös-Renyi graphs. It is clear that on linear functions (Table 2), the proposed method shows very strong results. In particular, our SID is the best in all cases. For SHD and F1, MERIT outperforms others for $p = 3, 5, 7$, while being comparable for $p = 10$. Patterns for uniform noise and logistic noise are quite similar. HCM mainly beats MERIT by SHD, while its SID is quite high for large p. This indicates that although HCM learns the skeleton correctly, the direction is often reversed. We then turn to the results on the MLPs (Table 3 in Appendix F). On the uniform noise, we again see that MERIT is

[1] `https://github.com/DAMO-DI-ML/AAAI2022-HCM` (MIT License)

Table 2: Results of RESIT, HCM and GS on the ER graphs with linear data in mean ± std. The first four columns use the uniform noise, and the others use the logistic noise.

Metric	Method	Uniform Noise				Logistic Noise			
		$p = 3$	$p = 5$	$p = 7$	$p = 10$	$p = 3$	$p = 5$	$p = 7$	$p = 10$
SHD	**MERIT**	**0.05±0.22**	**0.40±0.60**	**2.50±2.14**	6.25±3.11	**0.55±0.60**	**1.15±0.88**	**2.84±2.22**	6.63±3.56
	HCM	1.35±1.14	2.10±1.97	2.85±2.32	**4.80±2.12**	0.80±1.01	2.35±2.16	3.40±2.46	**5.42±2.19**
	GS	3.00±0.00	2.55±1.73	3.65±1.79	5.20±2.42	3.00±0.00	3.10±1.55	5.05±2.01	6.80±3.21
SID	**MERIT**	**0.05±0.22**	**0.55±1.00**	**4.35±3.95**	**12.20±7.75**	**1.00±1.34**	**1.50±1.32**	**5.26±4.29**	**13.74±10.53**
	HCM	3.50±2.31	6.45±5.84	9.30±8.44	16.75±9.12	2.15±2.48	6.65±5.57	10.65±9.50	18.58±10.34
	GS	6.00±0.00	8.10±4.77	11.00±6.57	19.30±12.59	5.90±0.45	9.20±4.93	14.00±7.41	25.30±12.77
F1	**MERIT**	**0.99±0.04**	**0.96±0.06**	**0.73±0.20**	0.60±0.22	**0.87±0.16**	**0.83±0.16**	**0.70±0.21**	0.55±0.25
	HCM	0.55±0.38	0.64±0.25	0.62±0.23	0.60±0.13	0.74±0.33	0.61±0.30	0.54±0.27	0.54±0.19
	GS	0.67±0.02	0.73±0.18	0.70±0.16	**0.72±0.09**	0.63±0.07	0.68±0.16	0.55±0.19	**0.63±0.15**

the best or competitive with the best method. On the logistic noise, MERIT performs best in SID while HCM performs best in SHD and F1 in most scenarios.

We also tested MERIT on Scale-Free graphs. The results are in Tables 4 and 5 in Appendix F. In general, it becomes more difficult for all methods to learn the SF graph. However, MERIT still outperforms competitors in many cases. For linear functions, MERIT is the best except for two cells in all metrics and dimensions. For MLPs, on the uniform noise, the best method depends on the dimension and the metric, while on the logistic noise, MERIT performs best on SHD and SID except for one cell. On F1, GS shows the best performance whereas MERIT ranks the second.

Real-world Datasets. We also compared MERIT with the SOTA method HCM on three real-world datasets `User knowledge data` [Kahraman et al., 2013] and `Algerian forest fire` [Abid and Izeboudjen, 2020] from the UCI repository [Kelly et al.], and `Pima Indians Diabetes Database` from Kaggle[2] to evaluate its practical performance. Details are provided in Appendix G. Across three datasets, MERIT consistently outperforms HCM in identifying causal relationships and preventing forbidden edges. Therefore, we believe MERIT's accuracy and reliability in causal discovery can make it a powerful tool for addressing real-world problems.

6 CONCLUSION

In this paper, we investigated causal discovery for mixed-type data, where both continuous and categorical variables are involved. We proposed a SCM with general function class for this problem. Specifically, each level of a categorical variable is modelled by an additive noise model. These values are put into the argmax function to determine the final outcome. Continuous variables follow the additive noise model. We proved the bivariate and multivariate identifiability of our SCM under rather weak assumptions on the noise. In addition, we gave concrete non-identifiable function classes for specific noise distributions. To learn the mixed-type SCM, we generalized the regression-based method RESIT to MERIT. MERIT starts by regressing continuous variables on the predictors and then assessing the dependence between the residuals and the predictors. If none of them is independent, it concludes that the leaf node must be categorical. Consequently, it identifies the corresponding categorical block and removes it. To select the parents of the categorical variables, MERIT uses conditional independence tests. Empirical studies show that the proposed method has excellent performance across various function classes and noise distributions and substantially outperforms the SOTA method HCM and GS in many scenarios.

Limitations: First, we assume the relation between the discrete and continuous variables can be modeled through an argmax model. Extensions to other models such as ordinal regression could be possible but are non-trivial. Second, while our method is proved to be consistent for any dimension, in practice, it performs the best on low-dimensional settings as the original RESIT. This is because the performance of dHSIC, the joint independence test, deteriorates on high-dimensional data [Zhu et al., 2020]. Other potential measures including the distance covariance [Jin and Matteson, 2018] have the same issue. Future advances in joint independence tests may eventually address the problem, while our proposed method provides a flexible framework that can incorporate them for high-dimensional causal discovery. Finally, if the ground truth is a DAG with contiguous blocks of categorical variables, a functional constraint on the connections between these variables is required to identify the complete causal ordering.

[2] https://www.kaggle.com/

Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" programme (RY, TV, JR).

References

Faroudja Abid and Nouma Izeboudjen. Predicting forest fire in algeria using data mining techniques: Case study of the decision tree algorithm. In Mostafa Ezziyyani, editor, *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, pages 363–370, Cham, 2020. Springer International Publishing. ISBN 978-3-030-36674-2.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Bryan Andrews, Joseph Ramsey, and Gregory F Cooper. Scoring bayesian networks of mixed variables. *International journal of data science and analytics*, 6:3–18, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, pages 2526–2556, 2014.

Wray Buntine. Theory refinement on bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 52–60. Elsevier, 1991.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1): 3741–3782, 2014.

Ruifei Cui, Perry Groot, and Tom Heskes. Copula pc algorithm for causal discovery from mixed data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16*, pages 377–392. Springer, 2016.

Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Takashi Ikeuchi, Mayumi Ide, Yan Zeng, Takashi Nicholas Maeda, and Shohei Shimizu. Python package for causal discovery based on lingam. *Journal of Machine Learning Research*, 24(14):1–8, 2023. URL http://jmlr.org/papers/v24/21-0321.html.

Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR, 2023.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10): 5168–5194, 2010.

Ze Jin and David S Matteson. Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete v-statistics. *Journal of Multivariate Analysis*, 168:304–322, 2018.

H Tolga Kahraman, Seref Sagiroglu, and Ilhami Colak. The development of intuitive knowledge classifier and the modeling of domain dependent data. *Knowledge-Based Systems*, 37:283–295, 2013.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository. URL https://archive.ics.uci.edu.

Murat Kocaoglu, Alexandros Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Yan Li, Rui Xia, Chunchen Liu, and Liang Sun. A hybrid causal structure learning algorithm for mixed-type data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7435–7443, 2022.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Wenqin Liu, Biwei Huang, Erdun Gao, Qiuhong Ke, Howard Bondell, and Mingming Gong. Causal discovery with mixed linear and nonlinear additive noise models: A scalable approach. In *Causal Learning and Reasoning*, pages 1237–1263. PMLR, 2024.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021.

Robert E McCulloch, Nicholas G Polson, and Peter E Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, 99(1):173–193, 2000.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, 2020.

Yang Ni and Bani Mallick. Ordinal causal discovery. In *Uncertainty in Artificial Intelligence*, pages 1530–1540. PMLR, 2022.

Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural Computation*, 27(3):771–799, 2015.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450, 2011.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

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

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):5–31, 2018.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3:121–129, 2017.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

Marco Scutari. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Andrew J Sedgewick, Ivy Shi, Rory M Donovan, and Panayiotis V Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, 17: 307–318, 2016.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110 (512):1726–1734, 2015.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Yan Zeng, Shohei Shimizu, Hidetoshi Matsui, and Fuchun Sun. Causal discovery for linear mixed data. In *Conference on Causal Learning and Reasoning*, pages 994–1009. PMLR, 2022.

K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 647. AUAI Press, 2009.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Changbo Zhu, Xianyang Zhang, Shun Yao, and Xiaofeng Shao. Distance-based and rkhs-based dependence metrics in high dimension. *The Annals of Statistics*, 48(6):3366–3394, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2 to 4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Appendix A.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See Appendix A.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 3 and Appendix B.
   (b) Complete proofs of all theoretical results. [Yes] See Section 3 and Appendix D.
   (c) Clear explanations of any assumptions. [Yes] See Sections 3 and 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Section 5 and appendix A
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 5.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Section 5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] See Section 5 and appendix G
   (b) The license information of the assets, if applicable. [Yes] See Section 5
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] See Appendix A.
   (d) Information about consent from data providers/curators. [Not Applicable] We use public datasets.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Causal discovery in mixed additive noise models: Supplementary Materials

A CODE AND COMPUTATION

The experiments were conducted on Intel Xeon Gold 6140 CPUs with 8 cores, 2.40 GHz frequency, and 5GB of memory for each core. The wall time is 24 hours. The MERIT and the data generation code are available at https://github.com/STAN-UAntwerp/MERIT.

The complexity is similar to the original RESIT as it performs $\mathcal{O}(p_1)$ ($p_1(p_2)$, number of continuous(categorical) variables) independence tests when measuring the dependence between the variables and the residuals. When the continuous variable X only has K number of categorical descendants, the procedure is repeated for $\mathcal{O}(p_2^K)$ to identify X and the categorical block, which remains acceptable when $K \ll p_1 + p_2 = p$. Note that the p_1 and p_2 can be arbitrary.

B PRELIMINARIES

In this section, we review some key definitions and results that support the proofs in the main paper.

B.1 Results on the Identifiability of SCM

Here, we review the bivariate and multivariate identifiability results from Peters et al. [2012] on SCM discussed in Sections 2 and 3 of the main paper.

Definition 3. *(Bivariate Identifiable Set) Let X, Y be two variables in some domain $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$. Let $\mathcal{F}|_2$ be a class of bivariate functions taking in one variable and the associated noise. We call a set $\mathcal{B} \subset \mathcal{F}|_2 \times \mathcal{P}_{\mathbb{R}} \times \mathcal{P}_{\mathbb{R}}$ containing combinations of functions $f \in \mathcal{F}|_2$ and distributions $\mathbb{P}_X, \mathbb{P}_{N_Y}$ of input X and noise N_Y bivariate identifiable in $\mathcal{F}$ if*

$$\begin{aligned} &(f, \mathbb{P}_X, \mathbb{P}_{N_Y}) \in \mathcal{B} \text{ and } Y = f(X, N_Y), N_Y \perp\!\!\!\perp X \\ &\Rightarrow \nexists\, g \in \mathcal{F}|_2 : X = g(Y, N_X), N_X \perp\!\!\!\perp Y \end{aligned} \tag{7}$$

holds. Additionally we require $f(X, N_Y) \not\perp\!\!\!\perp X$ for all elements in $\mathcal{B}$ with $N_Y \perp\!\!\!\perp X$.

In the multivariate case, we should require that the bivariate restricted function class is identifiable.

Definition 4. *(IFMOC) Let $\mathcal{B}$ be a bivariate identifiable set w.r.t. some $\mathcal{F}|_2$. We call a SCM with function class $\mathcal{F}$ a $(\mathcal{B}, \mathcal{F})$-IFMOC, if for all its functional models $X_i = f_i(\mathbf{PA}_i, N_i), i \in V$ for all $i \in V$, $j \in \mathbf{PA}_i$ and for all $x_{\mathbf{PA}_i \backslash \{j\}}$, we have*

$$f_i(x_{\mathbf{PA}_i\backslash\{j\}}, \underbrace{\cdot}_{X_j}, \underbrace{\cdot}_{N_i}) \in \mathcal{F}|_2. \tag{8}$$

Additionally, for all sets $S \subset V$ with $\mathbf{PA}_i \backslash \{j\} \subset S \subset \mathbf{ND}_i \backslash \{i, j\}$ ($\mathbf{ND}$ for non-descendants), there exists some x_S with positive probability such that the restricted bivariate identifiability holds

$$\left(f_i(x_{\mathbf{PA}_i\backslash\{j\}}, \underbrace{\cdot}_{X_j}, \underbrace{\cdot}_{N_i}), \mathbb{P}_{X_j|X_S = x_S}, \mathbb{P}_{N_i} \right) \in \mathcal{B}. \tag{9}$$

Then, we have the following results for multivariate identifiability.

Theorem 4 (Theorem 2 of Peters et al. [2012])**.** *If the distribution of $\{X_1, \dots, X_p\}$ is generated by a $(\mathcal{B}, \mathcal{F})$-IFMOC, then the causal DAG is identifiable.*

B.2 Results on Causal Discovery with Unobserved Variables

In the following, we review some recent results on causal discovery with unobserved variables from Maeda and Shimizu [2021]. Although we assume that all the variables in our model are observed (causal sufficiency), we need to select the suitable $B_Y = Re_Y \setminus S_Y$ in Algorithm 1 to remove so that the correct X_{k^*} can be identified in lines 10-13. In this procedure, we need to prevent the case when some categorical ancestors of X_{k^*} are mistakenly dropped (thus, not observed in the independence tests and regression), which leads to wrong results.

To characterize the undesired cases, we introduce the *unobserved causal path* and *unobserved backdoor path*, which would be eventually prevented by Algorithm 1 and the assumptions in the main paper.

Definition 5 (Definition 1 of Maeda and Shimizu [2021])**.** *A directed causal path from one observed variable X_i to another X_j is called an unobserved causal path (UCP) if it ends with the directed edge connecting X_j and some unobserved Y_k, i.e. $X_i \to \dots \to Y_k \to X_i$.*

Definition 6 (Definition 2 of Maeda and Shimizu [2021])**.** *A path between X_i and X_j is called a backdoor path if X_i and X_j have a common ancestor V_k. A backdoor path is called an unobserved backdoor path (UBP) if $\exists Y_m, Y_n$ unobserved such that $X_i \leftarrow \dots \leftarrow Y_m \leftarrow \dots \leftarrow V_k \to \dots \to Y_n \dots \to X_j$. In particular, $X_i \leftarrow Y_k \to X_j$ is a UBP.*

UCP and UBP in *Causal Additive Models* [Bühlmann et al., 2014] with unobserved variables (CAM-UV) can be fully characterized by the following Lemma.

Lemma 1 (Lemma 1 of Maeda and Shimizu [2021], informal expression)**.** *Let X be the set of the observed variables. Under the CAM-UV and certain assumptions, there is a UCP or UBP between the observed X_i and X_j if and only if*

$$\forall G_i, G_j (\textit{the regression function}), M_i \subset \{X \setminus \{X_i\}\}, M_j \subset \{X \setminus \{X_j\}\} \Longrightarrow (X_i - G_i(M_i)) \not\perp\!\!\!\perp (X_j - G_j(M_j)). \quad (10)$$

C COUNTERFACTUAL QUERIES FOR THE ARGMAX MODEL

In this section, we discuss learning the argmax model (Equation (2)) and answering counterfactual queries. In fact, the functions f_{k_y} in Equation (2) can be learned in some situations. For example, if $N_{k_y} \sim \text{Gumbel}(\mu, 1)$, then it is a multinomial logistic regression model (Ss proof of Corollary 2). If $N_{k_y} \sim \mathcal{N}(0, \sigma^2)$, then it is a multinomial probit model, which can be learned by the method in McCulloch et al. [2000].

Now the question is, whether we can gain any information on the noise N_{k_y} to answer counterfactual queries. The conclusion is:

The argmax model can answer some counterfactual queries but unfortunately not all.

We can consider a model with $K = 3$, $X \in \mathbb{R}$, and $Y = \arg\max(f_{k_1}(X) + N_{k_1}, f_{k_2}(X) + N_{k_2}, f_{k_3}(X) + N_{k_3})$. Suppose for a given $X = x_1$, the observed outcome is $Y = 3$. As we discussed above, we can assume that these f_{k_y} have been already estimated. Then, we know that $N_{k_2} < f_{k_3}(x_1) - f_{k_2}(x_1) + N_{k_3}$, $N_{k_1} < f_{k_3}(x_1) - f_{k_1}(x_1) + N_{k_3}$ (when N_{k_y} are continuous, the probability to be equal is zero). Equivalently, we can assume without loss of generality that N_{k_3} were 0, as substracting a constant value for each element in the model does not affect the outcome. Then we have: 1. N_{k_3} were 0. 2. N_{k_2} were less than $f_{k_3}(x_1) - f_{k_2}(x_1)$. 3. N_{k_1} were less than $f_{k_3}(x_1) - f_{k_1}(x_1)$. These upper bounds can be explicitly computed. Now we would like to answer the question, what would Y have been had X been set to another value x_2. Certainly, we can evaluate $(f_{k_1}(x_2), f_{k_2}(x_2), f_{k_3}(x_2))$, and we have the following answers to the query:

1. If $f_{k_3}(x_2) - f_{k_2}(x_2) > f_{k_3}(x_1) - f_{k_2}(x_1)$ and $f_{k_3}(x_2) - f_{k_1}(x_2) > f_{k_3}(x_1) - f_{k_1}(x_1)$, then the counterfactual $Y(x_2)$ would be 3.
2. If $f_{k_3}(x_2) - f_{k_2}(x_2) > f_{k_3}(x_1) - f_{k_2}(x_1)$ and $f_{k_3}(x_2) - f_{k_1}(x_2) < f_{k_3}(x_1) - f_{k_1}(x_1)$ (or $f_{k_3}(x_2) - f_{k_2}(x_2) < f_{k_3}(x_1) - f_{k_2}(x_1)$ and $f_{k_3}(x_2) - f_{k_1}(x_2) > f_{k_3}(x_1) - f_{k_1}(x_1)$), then the counterfactual could be 1 or 3 (2 or 3).

3. If $f_{k_3}(x_2) - f_{k_2}(x_2) < f_{k_3}(x_1) - f_{k_2}(x_1)$ and $f_{k_3}(x_2) - f_{k_1}(x_2) < f_{k_3}(x_1) - f_{k_1}(x_1)$, then the counterfactual $Y(x_2)$ could be 1, 2, or 3.

D PROOFS

Here we provide the proof of the theorems in Sections 3 and 4 of the main paper.

D.1 Theorem 1

Proof. We can factorize the joint distribution by

$$\begin{aligned} &P(X = k_x, Y = k_y) = P(X = k_x|Y = k_y)P(Y = k_y) \\ =&P'(Y' = k_y|X' = k_x)P'(X' = k_x) = P'(X' = k_x, Y' = k_y). \end{aligned}$$

Consider 4 equations with different input values, such that (1) (k_x, k_y) (2) (k_x, i_y), (3) (i_x, k_y), and (4) (i_x, i_y). By calculating $((1) \times (4))/((2) \times (3))$, we have the first equation. For the explicit formula, let $\Delta f_{k'_y}(k_x) := f_{k_y}(k_x) - f_{k'_y}(k_x)$, we have that

$$\begin{aligned} &P[Y = k_y|X = k_x] = P[k_y = \arg\max_{k'_y}(f_{k'_y}(k_x) + N_{k'_y})] \\ &= P[f_{k_y}(k_x) - f_{k'_y}(k_x) > N_{k'_y} - N_{k_y},\ \forall k'_y \neq k_y] \\ &= \int_{\mathbb{R}} P[N_{k_y} = z, N_{k'_y} < \Delta f_{k'_y}(k_x) + z, \forall k'_y \neq k_y]dz \\ &= \int_{\mathbb{R}} P[N_{k_y} = z]P[N_{k'_y} < \Delta f_{k'_y}(k_x) + z, \forall k'_y \neq k_y|N_{k_y} = z]dz \\ &= \int_{\mathbb{R}} P[N_{k_y} = z] \prod_{k'_y \neq k_y} P[N_{k'_y} < \Delta f_{k'_y}(k_x) + z|N_{k_y} = z]dz \\ &= \int_{\mathbb{R}} P[N_{k_y} = z] \prod_{k'_y \neq k_y} P[N_{k'_y} < \Delta f_{k'_y}(k_x) + z]dz \\ &= \int_{\mathbb{R}} p_{N_{k_y}}(z)dz \prod_{k'_y \neq k_y} \int_{-\infty}^{\Delta f_{k'_y}(k_x)+z} p_{N_{k'_y}}(z')dz' \end{aligned}$$

□

D.2 Corollary 1

Proof. We first prove the case when the noise distribution is Gumbel. The transition probability from X to Y is (see the proof of Corollary 2) with $P(Y = 1|X) = \frac{1}{1+\exp(-c-\beta_x X)}$. If a model $Y' \to X'$ also exists, then it means that $P(X' = 1|Y') = \frac{1}{1+\exp(-c-\beta_y Y')}$. Now, we let $(k_x, k_y, i_x, i_y) = (0, 0, 1, 1)$ and plug $P(Y = 1|X)$ in the left of Equation (4), we have that it equals to $\frac{\exp(-c)}{\exp(-c-\beta_x)} = \exp(\beta_x)$. Similarly, the right equals to $\exp(\beta_y)$. Therefore, to falsify Equation (4) and obtain identifiability, we need to show that $\beta_x \neq \beta_y$ under our assumption.
We prove by contradiction. Suppose such model exists and $\beta_x = \beta_y$, then the transition probability from X to Y and Y to X can be the same as the intercept c is equal (known) in both models. This means that there exists a Markov matrix $A = \begin{bmatrix} a & b \\ 1-a & 1-b \end{bmatrix}$, $0 < a, b < 1$ (as $c \in \mathbb{R}$) which satisfies

$$A^2 \begin{bmatrix} p_0 \\ 1 - p_0 \end{bmatrix} = \begin{bmatrix} p_0 \\ 1 - p_0 \end{bmatrix}, \text{ i.e. } (I - A^2) \begin{bmatrix} p_0 \\ 1 - p_0 \end{bmatrix} = 0, \tag{11}$$

by denoting $p_0 := P(X = 0)$. Now we calculate $I - A^2 = \begin{bmatrix} 1 - a^2 - b(1-a) & -ab - b(1-b) \\ -1 + a^2 + b(1-a) & +ab + b(1-b) \end{bmatrix}$ and find that it has rank 1. To satisfy Equation (11), it is thus sufficient to consider the first row, which is $(b - a - 1)(a - 1)p_0 +$

$(b-a-1)b(1-p_0)=0$. By taking out the common factor (can not be zero by the range of a,b), we end up with $ap_0+b(1-p_0)=p_0$. But this means that $A[p_0,1-p_0]^\top=[p_0,1-p_0]^\top$, and therefore X and Y have the same marginal distribution, which is a contradiction with our assumption. Therefore the model is identifiable.

Now we consider the general case. Although Equation (4) in this case does not have an analytic formula for the probability, we can still analyze it qualitatively. Let (k_x,k_y,i_x,i_y) be the same as above. The key observation on the left side (after permuting X and Y by the Bayes formula) is that $\frac{P(Y=k_y|X=k_x)}{P(Y=i_y|X=k_x)}$ is a function only on c (not β_x) as $k_x=0$, and $\frac{P(Y=i_y|X=i_x)}{P(Y=k_y|X=i_x)}$ is monotone in β_x due to the argmax model. The right-hand side also has the same expressions and functions on c and β_y. Therefore, after removing the common factor depending on c, Equation (4) will require that $\beta_x=\beta_y$ using the monotonicity. Then the proof proceeds in the same way as the above by discussing the property of the Markov matrix. □

D.3 Theorem 2

Proof. We prove by contradiction. That is, we first assume the distribution can be described by both models and then deduce the unique function class that f_k should belong to. Let $\Delta f_{k'_y}(x)=f_{k_y}(x)-f_{k'}(x)$. By definition, we have

$$\begin{aligned}
P[Y=k_y|X=x] &= P[k_y=\underset{1\le k'_y\le K_y}{\arg\max}(f_{k'_y}(x)+N_{k_y})]\\
&= P[f_{k_y}(x)-f_{k'_y}(x)>N_{k'_y}-N_{k_y},\ \forall k'_y\neq k_y]\\
&= \int_{\mathbb{R}} p_{N_{k_y}}(z)dz\prod_{k'_y\neq k_y}\int_{-\infty}^{\Delta f_{k'_y}(x)+z} p_{N_{k'_y}}(z')dz'\\
&= \int_{\mathbb{R}} p_{N_{k_y}}(z)\prod_{k'_y\neq k_y}F(\Delta f_{k'_y}(x)+z)dz
\end{aligned}$$

Let $p_{k_y}=P(Y=k_y)$ be the marginal distribution of Y (so as Y' under P' because $P(X,Y)=P'(X',Y')$). We have,

$$P'[X'=x]=\sum_{k'_y}P'[X'=x|Y'=k_y]P[Y'=k'_y]=\sum_{k'_y}p_{k'_y}P_{N_{X'}}(x-c_{k'_y}),$$

and thus by the Bayes formula,

$$P'[Y'=k_y|X'=x]=\frac{P'(X'=x,Y'=k_y)}{P'[X'=x]}=\frac{p_{k_y}P_{N_{X'}}(x-c_{k_y})}{\sum_{k'_y}p_{k'_y}P_{N_{X'}}(x-c_{k'_y})}$$

If the distributions generated by two causal models are the same, then we must have $P[Y=k_y|X=x]=P'[Y'=k_y|X'=x]$, and therefore the functions f_{k_y} and the noise distribution $p_{N'_x}$ must satisfy

$$\int_{\mathbb{R}} p_{N_{k_y}}(z)\prod_{k'_y\neq k_y}F(\Delta f_{k'_y}(x)+z)dz=\frac{p_{k_y}P_{N_{X'}}(x-c_{k_y})}{\sum_{k'_y}p_{k'_y}P_{N_{X'}}(x-c_{k'_y})}. \tag{12}$$

□

D.4 Corollary 2

Proof. In particular, if N_{k_y} follows the Gumbel noise, we have by the proof of Theorem 1 that

$$\begin{aligned} P[Y = k_y | X = x] &= \int_{\mathbb{R}} p_{N_{k_y}}(z) dz \prod_{k'_y \neq k_y} F_{k_y}(\Delta f_{k'_y}(x) + z) \\ &= \int_{\mathbb{R}} \exp\left(-z + e^{-z}\right) \prod_{k'_y \neq k_y} \exp\left(-e^{-z} e^{-\Delta f_{k'_y}(x)}\right) dz \\ &= \int_{\mathbb{R}} e^{-z} \exp\left(-e^{-z}\left(\sum_{k'_y \neq k_y} e^{-\Delta f_{k'_y}(x)} + 1\right)\right) dz \\ &= \int_0^{\infty} \exp\left(-t\left(\sum_{k'_y \neq k_y} e^{-\Delta f_{k'_y}(x)} + 1\right)\right) dt \\ &= \frac{1}{\sum_{k'_y \neq k_y} e^{-\Delta f_{k'_y}(x)} + 1} \\ &= \frac{e^{f_{k_y}(x)}}{\sum_{k_y=1}^{K_y} e^{f_{k_y}(x)}}. \end{aligned}$$

We can thus apply the above result to Equation (12) for level 1 and k_y ($\forall k_y$) respectively, and take the ratio between them. This would give us

$$\frac{e^{f_{k_y}(x)}}{e^{f_1(x)}} = \frac{p_{k_y} P_{N_{X'}}(x - c_{k_y})}{p_1 P_{N_{X'}}(x - c_1)}.$$

Since f_1 can be assumed to be 0 without loss of generality, we have by taking logarithm that

$$f_{k_y}(x) = \log\left(\frac{p_k}{p_1}\right) + \log\left(\frac{p_{N_{X'}}(x - c_{k_y})}{p_{N_{X'}}(x - c_1)}\right).$$

When we further assume $N_{X'}$ is the Gaussian noise, then $X' \sim \mathcal{N}(\mu_{k_y}, \sigma_{X'})$ given $Y' = k_y$. If we do the same as above for level 1 and k_y, then the right hand side after taking the ratio turns to

$$\frac{p_{k_y}}{p_1} \exp\left(-\frac{2x(\mu_1 - \mu_{k_y}) - \mu_1^2 + \mu_{k_y}^2}{2\sigma_{X'}^2}\right),$$

and therefore,

$$f_{k_y}(x) = \log\left(\frac{p_{k_y}}{p_1}\right) + \frac{2x(\mu_1 - \mu_{k_y}) - \mu_1^2 + \mu_{k_y}^2}{2\sigma_{X'}^2},$$

which is linear with respect to x.

□

D.5 Corollary 3

Proof. 1. Note that if all $f_{k_y} - f_{k'_y}$ are bounded, the left hand side of Equation (12) is always bounded in (0,1) because the distributions are assumed to have positive density on $\mathbb{R}$. If the condition for $P_{X'}$ holds and assuming that c_{k_y} is minimized at some k_y^*, then the limit of the right hand side as $x \to +\infty$ (without loss of generality) has to be 0 because there would be at least one $c_{k'_y} > c_{k_y^*}$ (otherwise c_{k_y} and thus $f'_{X'}$ is a constant, consequently there is no causation from Y to X, the proof completes). Then we check the limit of the left hand side. By using the dominate convergence theorem, we can put the limit inside the integral to deduce that,

$$\lim_{x \to +\infty} \int_{\mathbb{R}} p_{N_{k_y}}(z) \prod_{k'_y \neq k_y} F(\Delta f_{k'_y}(x) + z) dz = \int_{\mathbb{R}} p_{N_{k_y}}(z) \lim_{x \to +\infty} \prod_{k'_y \neq k_y} F(\Delta f_{k'_y}(x) + z) dz \tag{13}$$

and thus F must be equal to 0, $P_{N_{k_y}}-$ almost surely. Since $P(N_{k_y})$ is assumed to have density on $\mathbb{R}$, and F has density with support $\mathbb{R}$, this is not possible.

2. For the other condition, the reasoning is essentially the same. Without loss of generality, we assume such f_{k_y} and $f_{k'_y}$ satisfy $f_{k_y}(x) - f_{k'_y}(x) \to -\infty$. By dominant convergence, Equation (13) holds. Then, if a backward model exists, i.e. Equation (12) holds, the left-hand side has to be 0 because $\lim_{x\to\infty} F(f_{k_y}(x) - f_{k'_y}(x) + z) = 0$ for all z. This contradicts the asymptotic distribution of the noise on right hand side.

□

D.6 Theorem 3

We follow the notations in Section 4 of the main paper. The key to the justification of Algorithm 1 is to show that only the leaf node will be removed in each iteration. Therefore, the true causal order is identified by iteratively applying the procedure to the subgraphs (induced by line 17). In particular, a block of categorical variables may be removed before some continuous variable becomes the leaf node in the subgraph. More specifically, we need to show in each iteration:

1. (Necessity) If $X_i, i \in \{1, \ldots, p_1\}$ is the *ancestor* of a set of categorical variables $B_Y = \{Y_i\}_{i\subset\{p_1+1,\ldots,p_1+p_2\}}$ but is *not the ancestor* of any other continuous variables in the (sub)graph, then regressing X_{k^*} on the remaining variables other than B_Y leads to independent residuals. Therefore, X_i and B_Y may be identified in lines (11-13). In particular, when several choices of X_i and B_Y lead to independent residuals, Algorithm 1 chooses the one with the least number of categorical descendants. This avoids removing the categorical ancestors of X_i when it is the leaf node in the graph.
2. (Sufficiency) If X_i is the ancestor of some X_j or Y_j, then lines (11-13) will not be executed. That is X_i will not be regarded as a leaf node, no matter what S_Y is in the consideration.

We prove them by induction. For clarity, we assume that $\mathcal{G}_0$ is the current graph on $S_0 := S \cup Re_Y$, and $\mathcal{G}_1$ is the subgraph induced by $S_1 := S \cup S_Y$ for which X_i is the continuous variable under testing. With a slight abuse of notation, we also use S_0 and S_1 to denote the set of variables. Note that any UCP or UBP in $\mathcal{G}_1$ is a path in $\mathcal{G}_0$.

Lemma 2. *Suppose there is a directed path with only the observed variables from X_i to its youngest child X_j in $\mathcal{G}_0$ and there is no UBP or UCP between them or X_i and $\mathbf{DE}(X_j)$. Then X_i will not be identified as the leaf node in $\mathcal{G}_0$.*

Lemma 3. *Suppose there is a UBP or UCP from X_i to X_j in $\mathcal{G}_1$, then X_i will not be identified as the leaf node in $\mathcal{G}_0$*

Lemma 4. *Suppose there exists a categorical variable $Y_j \in \mathcal{G}_1$ such that, there is a directed path, UBP or UCP from X_i to Y_j in $\mathcal{G}_1$, then X_i will not be identified as the leaf node in $\mathcal{G}_0$.*

The cases in Lemma 2, 3 and 4 are illustrated in the following graphs.

Proof of Lemma 2. Here we assume X_j is the youngest child of X_i in $\mathcal{G}_0$. If such X_j does not exist (e.g. the youngest child is categorical and is observed in $\mathcal{G}_1$, or there exists a UCP or UBP between X_i and X_j in $\mathcal{G}_1$), then it fits into the other two lemmas. Next, we show that X_i will not lead to independent residuals in $\mathcal{G}_1$ by contradiction.

We can consider a subgraph $\mathcal{G}'_0$ of $\mathcal{G}_0$ by removing the descendants (DE) of X_j. Clearly, X_j is the leaf node in $\mathcal{G}'_0$, thus regressing X_j on the remaining variables in $\mathcal{G}'_0$ leads to independent residuals.

On the other hand, X_i leads to independent residuals in $\mathcal{G}_1$, therefore we assume $X_i = f(S_1 \setminus X_i) + \epsilon_i$ with $\epsilon_i \perp\!\!\!\perp (S_1 \setminus X_i)$. We then have $\mathbf{DE}(X_j)$ is independent of X_i given $S_1 \setminus (X_i \cup \mathbf{DE}(X_j))$ (from the Markov property of $\mathcal{G}_0$ and the conditions), which means that they can be removed in $\mathcal{G}_1$ to get $\mathcal{G}'_1$. Now, let $S'_1 = S_1 \setminus (\{X_i, X_j\} \cup \mathbf{DE}(X_j))$, $S_{\text{add}} = \{Y \in S_0 \text{ categorical}, Y \in \mathbf{PA}(X_j) \text{ in } \mathcal{G}_0\} \setminus S_1$, we define $S_{\text{cond}} = S'_1 \cup S_{\text{add}}$. Clearly $S_{\text{add}} \perp\!\!\!\perp X_i | S'_1$ and S_{cond} includes all parents of X_j because there is no UBP,UCP between X_i and X_j. Therefore, we conclude that

$$X_i|_{S_{\text{cond}}} = f(X_j|_{S_{\text{cond}}}, S_{\text{cond}}) + \epsilon_i, \quad X_j|_{S_{\text{cond}}} = g(X_i|_{S_{\text{cond}}}, S_{\text{cond}}) + \epsilon_j,$$

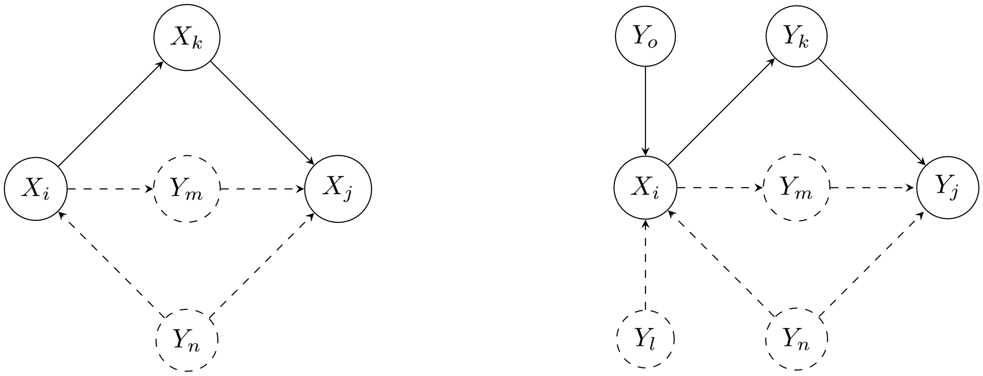

Figure 1: The dashed nodes and the dashed edges are unobserved in $\mathcal{G}_1$, but is observed in $\mathcal{G}_0$. $X_i \to Y_m \to X_j$, and $X_i \to Y_m \to Y_j$ are UCPs, while $Y_n \to X_i, Y_n \to X_j$ and $Y_n \to X_i, Y_n \to Y_j$ are UBPs. In particular, there is no UCP or UBP regarding X_i, X_k and X_i, Y_k.

with $\epsilon_{i,j}$ independent of the predictors respectively, which contradicts the restrictive bivariate identifiability of the continuous variables. □

Proof of Lemma 3. For this lemma, we first prove a similar result to Lemma 1 under our Assumptions 1 and 3. Let W_i be the additive effects of the unobserved categorical parents of X_i (due to S_Y), and W_j the same for X_j. If there is a UCP between X_i and X_j, we have by condition 4 in Assumption 3 that $N_{X_i} \not\perp\!\!\!\perp W_j$ (note that when $X_i \in \mathbf{PA}(W_j)$, this can also follow from causal minimality). Therefore, $(W_i + N_{X_i}) \not\perp\!\!\!\perp (W_j + N_{X_j})$ as the external noise variables N_X are mutually independent. If there is only UBP between X_i and X_j, then according to the definition there exists a variable V (either continuous or categorical) such that $N_V \not\perp\!\!\!\perp Y_m$, and $N_V \not\perp\!\!\!\perp Y_n$, where $Y_{m,n}$ are unobserved (here we used the same notation for the noise in the categorical variables). Thus, by letting $M_i = \mathbf{PA}(X_i) \setminus \{Y|\text{components of } W_i\}$ and $M_j = \mathbf{PA}(X_j) \setminus \{Y|\text{components of } W_j\}$, $N = N_V$ and applying condition 3 in Assumption 3 as the proof in Maeda and Shimizu [2021], we have $W_i \not\perp\!\!\!\perp W_j$. This means that $(W_i + N_{X_i}) \not\perp\!\!\!\perp (W_j + N_{X_j})$ remains hold.

Now we let $G_k^*(M_k)$ $k = i, j$ be $G_k(M_k)$ (the regression function of X_k on M_k) minus the causal functions $f_k(\mathbf{PA}(X_k) \setminus (Re_Y \setminus S_Y))$ of X_k. Since the true causal function is additive with respect to the categorical variables (Assumption 3), Equation (10) is equivalent to $(W_i + N_i - G_i^*(M)) \not\perp\!\!\!\perp (W_j + N_j - G_j^*(M))$. Then we want to show that this is equivalent to $(W_i + N_{X_i}) \not\perp\!\!\!\perp (W_j + N_{X_j})$. This part follows exactly from the proof of Lemma 1 using condition 2 in Assumption 3, thus we omit it here.

To conclude, if there is a UCP or UBP between X_i and X_j in our MANM, then Equation (10) holds. Now, we let M_i be all the variables in $\mathcal{G}_1$, and $M_j = \phi$, then the equation simply translates to the fact that the residual of X_i regressing on the remaining variables can not be independent of X_j. Therefore, X_i will not be identified as the leaf node in line 10. □

Proof of Lemma 4. We split the proof into two cases.

Case 1: There exists an observed direct path from X_i to Y_j youngest child in $\mathcal{G}_0$, and there is no UBP or UCP in between. Using the same reasoning as the proof of Lemma 2, if X_i leads to independent residuals, we can identify a set S_{cond} in $\mathcal{G}_0$, such that

$$X_i|_{S_{\text{cond}}} = f_{k_y}(Y_j|_{S_{\text{cond}}}, S_{\text{cond}}) + \epsilon_i, \quad Y_j|_{S_{\text{cond}}} = \underset{k_y \in \{1,\ldots,K_y\}}{\arg\max} \left(f_{k_y}(X_i|_{S_{\text{cond}}}, S_{\text{cond}}) + \epsilon_j\right).$$

But this contradicts the restricted bivariate identifiability between the continuous and the categorical variables in the MANM.

Case 2: The first situation is that there is a UCP between X_i and Y_j, through some observed X_k, Y_k, and unobserved Y_m. Since the proof is similar to Lemmas 1 and 3 on UCP by setting $M_j = \phi$, we omit it here.

Given all the known results, we can assume that X_i has no descendants that are continuous variables, therefore it will be the true leaf node of the subgraph if certain categorical variables are removed. Thus, we want to prove here that only if all the categorical descendants of X_i are removed, X_i leads to independent residuals. Here we consider a UBP or UCP between X_i and Y_j in $\mathcal{G}_1$. Without loss of generality, we may assume Y_j has no descendants in $\mathcal{G}_0$ observed in $\mathcal{G}_1$(otherwise, we choose that one as Y_j). Let $S_{Y_{un}} := Re_Y \setminus S_Y$ be the set of unobserved variables. We first assume that $S_{Y_{un}} \cap \mathbf{PA}(X_i) \neq \phi$. We have in $\mathcal{G}$, there exists a function f and ϵ independent of the components of f such that

$$X_i = f(S \setminus \{X_i\}, S_Y \setminus \mathbf{DE}(X_i), S_{Y_{un}} \cap \mathbf{PA}(X_i)) + \epsilon \tag{14}$$

On the other hand, since X_i leads to independent residuals in $\mathcal{G}_1$, we have a function g and ϵ independent of the components of g such that

$$X_i = g(S \setminus \{X_i\}, S_Y \setminus \mathbf{DE}(X_i), \mathbf{DE}(X_i) \cap S_Y) + \epsilon' \tag{15}$$

Due to the discussion at the beginning of the proof, we can assume $\mathbf{DE}(X_i)$ are all categorical. Now, we consider the conditional distribution of X_i given $(S \setminus \{X_i\}) \cup (S_Y \setminus \mathbf{DE}(X_i))$. By Equation (14), it is a mixture distribution in the form of $P(X) = \sum_{i=1}^{K} c_i P_\epsilon(X, \theta_{c_i})$, $\sum_{i=1}^{K} c_i = 1, K > 1$, and θ_{c_i} is the parameter to indicate the location shifts, see the explanation in Section 2 of the main paper. According to Assumption 2, it can be uniquely identified, meaning that all c_i and P_ϵ are determined. Now, we list all the possible combinations of $S_{Y_{un}} \cap \mathbf{PA}(X_i)$ and $\mathbf{DE}(X_i) \cap S_Y$. Since the distribution is an identifiable mixture, there is a deterministic map to map every combination of $S_{Y_{un}} \cap \mathbf{PA}(X_i)$ and $\mathbf{DE}(X_i) \cap S_Y$ to (c_i, θ_{c_i}), and thus given $S \setminus \{X_i\}, S_Y \setminus \mathbf{DE}(X_i), S_{Y_{un}} \cap \mathbf{PA}(X_i)$ and $\mathbf{DE}(X_i) \cap S_Y \setminus \{Y_j\}$ (non-desendants of Y_j by design), the value of Y_j either has zero probability on some levels or can take values on all levels. For the former, since the noise N_{k_y} takes values on $\mathbb{R}$, it contradicts the argmax model. For the latter, it means that Y_j and X_i are independent given observed variables (Equation (15)), so as Y_j and the residual from regression. But this contradicts Assumption 3 which asserts that when the residual and Y_j are dependent with N_{X_i}, the residual and Y_j are also dependent. Then we conclude that the assumption of MANM is violated and the causal direction between the variables can not be determined, which is a contradiction.

It remains to deal with the case that $S_{Y_{un}} \cap \mathbf{PA}(X_i) = \phi$. In this case, all true parents of X_i are included in $\mathcal{G}_1$ and we may assume that $X_i \not\perp\!\!\!\perp (\mathbf{DE}(X_i) \cap S_Y) | (S \setminus \{X_i\}, S_Y \setminus \mathbf{DE}(X_i))$ (so g is not a constant on $\mathbf{DE}(X_i) \cap S_Y$ given other variables), otherwise, some variables can be removed by the causal minimality. But this means that $X_i | (S \setminus \{X_i\}, S_Y \setminus \mathbf{DE}(X_i))$ is a mixture distribution which contradicts the Assumption 2. □

Proof of Theorem 3. We now prove Theorem 3. We start with Algorithm 1, for which we need to show that if and only if X_i is the leaf node, then it will be selected in line 10 under proper S_Y.

(Necessity) We know that regressing a continuous variable X_i on the remaining variables $S_0 \setminus X_i$ in the graph yields residuals that are independent of $S_0 \setminus X_i$ if X_i is the leaf node. This follows from the fact that $\mathbb{E}[X_i | S_0 \setminus X_i] = \mathbb{E}[X_i | \mathbf{PA}(X_i)]$ and $X_i = \mathbb{E}[X_i | \mathbf{PA}(X_i)] + e_i$ where e_i is the independent residual. Similarly, when X_i is not an ancestor of any other continuous variables, the above holds if we replace S_0 by $S_0 \setminus \mathbf{DE}(X_i)$. This means that Algorithm 1 can identify X_i as the leaf node in the graph $\mathcal{G}_1$ where $S_Y = Re_Y \setminus \mathbf{DE}(X_i)$. Since line 4 starts with the largest subset, i.e. Re_Y, Algorithm 1 chooses the one with the least number of categorical descendants. Therefore, once a pair of $(X_i,\ S_Y)$ is found, it will not drop any categorical variables whose causal order is before X_i so that the resulting $\mathcal{G}_1$ includes all the variables necessary for identifying the causal order in the subsequent procedures.

(Sufficiency) We only need to show that Lemmas 2 to 4 already included all the possible cases of selecting a wrong X_i. In fact, when X_i is not a leaf node, it is either the parent of a continuous variable X_j or a categorical variable Y_j. For the first case, Lemmas 2 and 3 solve the issue. For the latter, Lemma 4 solves the issue.

The justification of the second phase in the original RESIT (which finds all the causal parents of the continuous variables) follows from Lemma 38 of Peters et al. [2014]. Therefore, we conclude that Algorithm 1 correctly identifies the causal order of the continuous variables and the categorical blocks. And the parents of the continuous variables are also correctly identified.

For Algorithm 2, without loss of generality, we first let $\mathbf{CH}_c(X_i), \mathbf{CH}_c(Y_j)$ be the known continuous children of X_i, Y_j. Note that we do not need to consider the known categorical children of X_i, as their order comes after Y_j, whereas the child of X_i and Y_j in the block B_Y s.t. $Y_j \in B_Y$ has not been identified yet. Now suppose Y_j has no directed path to any $Y \notin \mathbf{CH}_c(X_i)$ in the same block B_Y (see e.g. Figure 2), then we have $X_i \in \mathbf{PA}(Y_j)$ if and only if $X_i \not\perp\!\!\!\perp Y_j | (\mathbf{CH}_c(X_i) \setminus \mathbf{CH}_c(Y_j))$ (in the figure, only need to be conditioned on X_k to identify both Y_j and

Y_l) as long as the causal order learned in Algorithm 1 is correct. This is because there will be no child of X_i in B_Y that needs to be given.

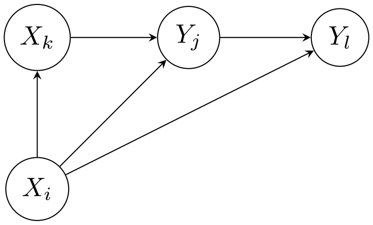

Figure 2: Y_j is only connected with $Y_l \in B_Y$ which is also the child of X_i.

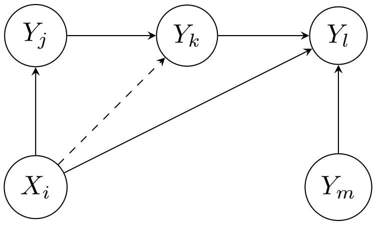

Figure 3: Y_l is a collider. The dashed edge means the connection may (not) exist in $\mathcal{G}_0$, which results in different treatments.

In more general cases where such connection exists (see e.g. Figure 3), we still need to distinguish Y_k from the others as for each $Y \in \{Y_j, Y_k, Y_l\}$,

$$X_i \not\perp\!\!\!\perp Y | (\mathbf{CH}_c(X_i) \setminus \mathbf{CH}_c(Y)), \tag{16}$$

due to $X_i \to Y_j \to Y_k$ (recall currently we assume $\mathbf{CH}_c$ only includes the known continuous children) but $X_i \to Y_k$ might not exist. Note that Y_m would not lead to conditional dependence and, therefore is already excluded. Now, the observation is that for any subset S_{-Y_l} of $\{Y_j, Y_k\}$, and any subset S_{-Y_j} of $\{Y_l, Y_k\}$,

$$X_i \not\perp\!\!\!\perp Y_l | (\mathbf{CH}_c(X_i) \setminus \mathbf{CH}_c(Y_l)) \cup S_{-Y_l}, \quad X_i \not\perp\!\!\!\perp Y_j | (\mathbf{CH}_c(X_i) \setminus \mathbf{CH}_c(Y_j)) \cup S_{-Y_j}, \tag{17}$$

however, when letting $S_{-Y_k} = \{Y_j\}$ (so the collider is removed), $X_i \perp\!\!\!\perp Y_k | (\mathbf{CH}_c(X_i) \setminus \mathbf{CH}_c(Y_k)) \cup S_{-Y_k}$, because Y_l is actually the child of X_i and Y_k, and once it is removed, there is no backdoor path. Same argument applies to the fully general case where several directed paths from X_i to Y_l exist. Therefore, among all Y such that Equation (16) is satisfied, we identify the variables for which Equation (17) holds for all S_{-Y} corresponding to Y. Then, Y is indeed the child of X_i, and is added to $\mathbf{CH}_c(X_i)$. The others are removed, which prevents Y_k form being identified as the child node. The last part of Algorithm 2, which identifies the remaining connections between the categorical variables, is justified by the standard algorithm and the identifiability property of MANM, which means only a certain function class is allowed to model the distribution. The proof is complete. □

Remark 2. *Here, we discuss several cases where Assumption 2 can be relaxed.*

The conditions in Assumption 2 are mainly used in the proof of Lemma 4 (Case 2). The identifiability of the mixture is to ensure that there will be a deterministic map from the space of the categorical parents to the space of the mixture parameters. When the mixture is not fully identifiable (e.g. several mixtures that induce the same distribution), the contradiction may still hold. For example, the following density can be represented by a mixture (M1) of uniform(-0.5,0.5) and uniform(0,1) with equal probability, or a mixture (M2) of uniform(-0.5,0), uniform(0,0.5), uniform(0.5, 1) with probability (1/4,1/2,1/4):

$$p(x) = \begin{cases} 1/4 & -0.5 < x < 0 \\ 1/2 & 0 < x < 0.5 \\ 1/4 & 0.5 < x < 1. \end{cases}$$

Now, suppose $Y_1 \in \{0,1\}$ is the only unobserved categorical parent in Equation (14) which contributes to the mixture M1 in the conditional distribution such that $Y_1 = 0 \Rightarrow x \in (-0.5, 0.5)$. (Y_2, Y_3) the "identified" parents (Y_2 corresponds to Y_j in the proof) in Equation (15) that contribute to the mixture M2 such that when the pair equal is (0,0) or (0,1), $x \in (-0.5, 0)$, and (1,0), (1,1) correspond to the other two piles. But then we see given all other variables, Y_2 can only be 1, and therefore still no $N_{k_{y_2}}$ involved, which is a contradiction. The second condition in Assumption 2 is used to illustrate that the descendants of X_i still need to be removed when $S_{Y_{un}} \cap \boldsymbol{PA}(X_i) = \phi$ and there exists UCP and UBP (fourth paragraph in the proof). However, using conditions 1 and 2 in Assumption 3, an independent residual is impossible when regressing Y_j on the parents of X_i. Therefore, this condition may not be needed.

E MERIT ON AN EXAMPLE

We consider a toy example on $(Y_1, X_1, Y_2, Y_3, X_2)$ (X_1,X_2 continuous, others categorical), with connections,$Y_1 \to X_1$, $X_1 \to Y_2$, $X_1 \to Y_3$, $Y_1 \to Y_3$, $Y_3 \to X_2$, and $Y_2 \to Y_3$ as shown in Figure 4. At the start of Algorithm 1, it sets $S_Y = Re_Y = \{Y_1, Y_2, Y_3\}$ (Line 4) and regresses X_1 on the remaining variables. X_2 will be identified as the first sink as the original RESIT. For the next step, since the children of X_1 are contained in the regressors, the residual would not be independent of X_1 (See Theorem 3). Therefore, it starts to remove the categorical variables, until the residual becomes independent. Specifically, it iterates on the subset of Re_Y (Line 4) with $S_Y = \{Y_2, Y_3\}, \{Y_1, Y_3\}, \{Y_1, Y_2\}, \ldots, \{Y_3\}, \phi$, and regresses X_1 on S_Y. By Theorem 3 and the searching order, if and only if $S_Y = \{Y_1\}$ (so the children of X_1 are correctly dropped), the residuals would be independent. Then, line 10 returns true for $S_Y = \{Y_1\}$ and lines 11-14 are executed. Specifically, $B_Y = Re_Y \backslash S_Y = \{Y_2, Y_3\}$ indicates the children/descendants of X_1. The remaining categorical variable Re_Y is updated to $\{Y_1\}$. X_1 is a potential parent of Y_1, Y_2. Y_1 is a potential parent of X_1. The learned causal order is then (Y_1, X_1, B_Y, X_2), while the order inside B_Y is not known. Algorithm 1 now ends.

Now, Algorithm 2 starts to find the continuous parents of Y_1, Y_2 (until line 7), and identifies the connections $X_1 \to Y_2, Y_3$. Finally, it remains to identify the connections between the categorical variables (line 8). By the PC algorithm or the BDeu, connections $Y_1 - Y_2$, $Y_2 - Y_3$ are detected. Since we have identified the causal order (Y_1, X_1, B_Y, X_2) in Algorithm 1, we can assert that $Y_1 \to Y_2$. Now for Y_2, Y_3, we can fit classification models (e.g. multinomial logistic regression, multinomial probit model) and compare the learned functions with the restrictions (e.g. zero intercept) to prefer the causal direction $Y_2 \to Y_3$, see Corollary 1.

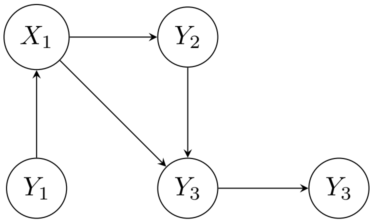

Figure 4: Causal relationships in the example.

Remark 3. *We highlight that requiring such functional and/or distributional constraints is not specific to our method. Identifiability results in causal discovery are usually based on restricting the function class (LiNGAM, ANM, post-nonlinear model), including the recent Ordinal causal discovery [Ni and Mallick, 2022] on categorical variables, where both functional and ordinal constraints are required. Alternatively, entropy constraints [Kocaoglu et al., 2017] have also been considered for categorical/mixed-type causal discovery. Naturally, causal discovery among categorical variables is more challenging as any Markov transition matrix can model the relationship. Therefore, more constraints and postulates are required for the identifiability. We also would like to emphasize that there can be many possibilities for such a restriction on the function class other than those we have described in Corollary 1, and there is no canonical choice that admits the widest application. Therefore in practice, it is also reasonable to use the subject-matter knowledge to infer, e.g. region to rain rather than rain to region as in our experiment on the* `Algerian forest fires` *dataset*

F FULL SIMULATION RESULTS

In this section, we provide the full results for the experiments in Section 5.

For linear data, MERIT remains the best method on SF graphs. For MLP data, MERIT is the best or remains competitive with the best method for ER graphs with uniform noise and SF graphs with logistic noise. HCM works best for ER graphs with logistic noise. The best method for SF graphs with uniform noise, however, depends on the dimension and the metric.

Table 3: Results of MERIT, HCM, and GS on the ER graphs with MLP data in mean ± std. The first four columns use the uniform noise, and the others use the logistic noise. MERIT is competitive with HCM for uniform noise, while the latter performs the best for the logistic noise. GS is relatively the worst.

Metric	Method	Uniform Noise				Logistic Noise			
		$p = 3$	$p = 5$	$p = 7$	$p = 10$	$p = 3$	$p = 5$	$p = 7$	$p = 10$
SHD	**MERIT**	**0.75±0.79**	1.53±1.23	**2.38±1.94**	5.50±4.09	**0.85±0.49**	1.50±1.29	2.07±1.38	5.13±3.14
	HCM	1.05±0.83	**1.35±1.32**	**2.38±1.33**	**2.83±2.64**	**0.85±0.37**	**1.14±0.77**	**1.86±1.99**	**2.75±1.58**
	GS	2.80±0.62	2.70±1.53	4.55±2.54	6.82±3.15	3.00±0.00	3.20±1.88	4.15±2.52	7.18±2.77
SID	**MERIT**	**0.85±0.99**	3.18±3.78	**5.61±5.17**	9.17±7.73	**0.90±0.64**	**2.21±2.64**	**3.79±3.53**	13.38±10.46
	HCM	1.75±1.89	**3.12±3.44**	6.15±4.34	**7.33±7.81**	1.30±0.73	2.29±2.49	4.43±5.39	**9.63±9.91**
	GS	5.60±1.23	7.20±5.22	10.45±7.19	18.00±13.67	6.00±0.00	9.35±6.67	11.00±7.97	21.29±16.06
F1	**MERIT**	**0.83±0.19**	0.73±0.28	**0.74±0.20**	0.51±0.29	**0.82±0.13**	0.71±0.28	0.74±0.23	0.52±0.24
	HCM	0.75±0.25	**0.79±0.22**	0.70±0.20	**0.74±0.23**	0.81±0.10	**0.76±0.19**	**0.82±0.18**	**0.77±0.13**
	GS	0.63±0.07	0.67±0.15	0.60±0.23	0.62±0.14	0.64±0.04	0.57±0.20	0.63±0.25	0.57±0.12

Table 4: Results of MERIT, HCM, and GS on the SF graphs with linear data in mean ± std. The first four columns use the uniform noise, and the others use the logistic noise. MERIT is the best method in almost all entries.

Metric	Method	Uniform Noise				Logistic Noise			
		$p = 3$	$p = 5$	$p = 7$	$p = 10$	$p = 3$	$p = 5$	$p = 7$	$p = 10$
SHD	**MERIT**	**0.15±0.49**	**2.06±1.39**	**3.29±1.94**	10.27±6.18	**0.35±0.49**	**2.38±1.54**	**4.79±2.55**	**7.91±3.83**
	HCM	0.80±0.83	3.20±2.14	7.95±4.24	9.90±5.04	1.20±1.20	3.15±1.73	7.00±4.50	9.20±5.83
	GS	3.00±0.00	5.70±1.38	7.45±2.31	**9.50±3.71**	3.00±0.00	5.90±1.17	8.25±1.89	10.50±3.76
SID	**MERIT**	**0.20±0.70**	**3.25±2.86**	**6.29±3.29**	**27.18±21.23**	**0.60±1.05**	**4.13±3.12**	**11.79±8.14**	**26.91±11.43**
	HCM	2.40±2.19	8.90±5.78	20.10±8.38	34.10±12.97	2.70±2.64	8.30±4.01	20.20±9.68	34.35±15.19
	GS	6.00±0.00	18.40±2.82	31.20±10.00	46.90±23.01	5.80±0.62	18.55±2.42	31.40±7.60	51.30±16.38
F1	**MERIT**	**0.96±0.12**	**0.81±0.13**	**0.81±0.10**	0.60±0.23	**0.92±0.11**	**0.76±0.16**	**0.68±0.21**	**0.65±0.19**
	HCM	0.74±0.28	0.63±0.23	0.51±0.19	0.58±0.12	0.62±0.40	0.63±0.20	0.54±0.21	0.59±0.13
	GS	0.67±0.00	0.65±0.07	0.61±0.12	**0.68±0.11**	0.63±0.08	0.60±0.12	0.58±0.09	0.61±0.11

Table 5: Results of MERIT, HCM, and GS on the SF graphs with MLP data in mean ± std. The first four columns use the uniform noise, and the others use the logistic noise. MERIT is either the best method or competitive with the best method for logistic noise on each metric. However, the best method for uniform noise depends on the dimension and the metric.

Metric	Method	Uniform Noise				Logistic Noise			
		$p = 3$	$p = 5$	$p = 7$	$p = 10$	$p = 3$	$p = 5$	$p = 7$	$p = 10$
SHD	**MERIT**	1.00±0.65	**2.64±1.15**	6.00±2.79	10.09±3.70	**0.70±0.86**	**3.00±0.94**	**5.00±1.73**	10.10±2.96
	HCM	**0.90±1.02**	2.86±1.10	**4.18±1.83**	**7.64±2.46**	0.95±1.00	**3.00±1.33**	5.54±1.61	10.70±2.41
	GS	2.80±0.62	3.05±1.99	5.00±2.83	8.29±2.59	2.80±0.62	3.25±1.97	5.10±2.29	**8.94±3.15**
SID	**MERIT**	**1.15±0.88**	**5.43±3.94**	15.36±10.03	30.64±16.10	**0.90±1.41**	**5.80±3.33**	**10.15±6.09**	**24.40±12.54**
	HCM	1.65±2.28	6.79±4.82	12.91±6.39	**24.27±11.66**	1.70±2.25	7.30±3.59	15.92±6.90	37.00±12.11
	GS	5.60±1.23	8.80±5.43	**12.35±7.53**	25.12±12.74	5.60±1.23	8.85±6.32	13.05±7.03	27.65±13.21
F1	**MERIT**	**0.77±0.17**	**0.71±0.17**	0.57±0.23	0.57±0.18	**0.82±0.25**	0.68±0.15	0.63±0.18	0.54±0.15
	HCM	0.74±0.34	0.67±0.17	**0.68±0.15**	0.63±0.14	0.73±0.33	0.62±0.19	0.55±0.16	0.44±0.13
	GS	0.62±0.07	0.69±0.24	**0.68±0.21**	**0.65±0.13**	0.64±0.07	**0.71±0.17**	**0.68±0.13**	**0.61±0.15**

G REAL-WORLD DATASETS

We now include illustrations of the proposed algorithm on three real datasets, `User knowledge modeling`, `Pima-diabetes`, and `Algerian forest fire`.

G.1 User Knowledge Modeling

In this Section, we compare MERIT with HCM on the real dataset `User knowledge modeling` from the UCI repository (https://archive.ics.uci.edu/dataset/257/user+knowledge+modeling). It is a classification task where the goal is to measure the knowledge level (discrete) of students (UNS) using predictors: STG (The study time for goal object), SCG (The repetition number for goal object), STR (The study time of user for objects related with goal object), LPR (The exam performance of user for related objects with goal object), PEG (The exam performance of user for goal objects). The target UNS is calculated based on the features using the expert's opinion. Therefore, the target variable is supposed to come last in the causal order (i.e., it should have no child nodes in the causal graph).

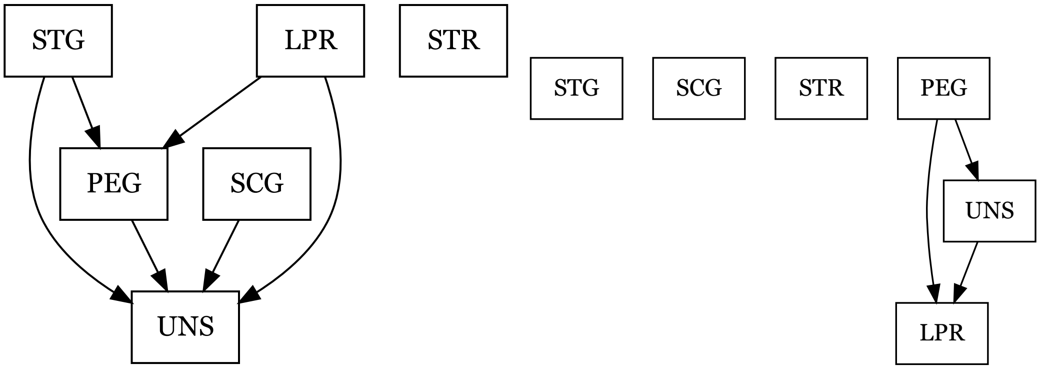

Figure 5: Left: Causal graph estimated by MERIT. Right: Causal graph estimated by HCM.

As we can see from the causal graph of MERIT, it correctly identifies 4 causes of UNS out of 5. In addition, it finds that the study time for goal object (STG) is the cause of the exam performance of user for goal objects (PEG). This is as expected. The edge between LPR (The exam performance of user for related objects) and PEG might be less expected. It is more likely to be a correlation rather than a causation. But given the relatively small size of the dataset, such inaccuracies can occur.

We contrast our result with the causal graph obtained by the HCM. HCM identifies far fewer edges. For the target variable, it only finds the edge from PEG to UNS. In addition to that, there is an edge from UNS to LPR which is reversed (because UNS is calculated on LPR by the expert's knowledge). Finally, it also finds the relation between LPR and PEG, which is more likely to be a vanilla correlation.

To conclude, MERIT shows better performance on this dataset than HCM in the sense that: 1. It finds more causal parents of the discrete target. 2. It identifies a causation between two predictors which is reasonable.

G.2 Pima Indians Diabetes Database

In this Section, we compare MERIT and HCM on the real dataset `Pima Indians Diabetes Database` [Smith et al., 1988] (fetched from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database). This is a dataset for classification, where the target “Outcome" is binary which indicates whether the person in query has diabetes. The predictors are: Pregnancies, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, Diabetes_Pedigree_Function, and Age. Among them, the BMI and Diabetes_Pedigree_Function are continuous, others are ordinal or discrete. Two predictors Glucose and Blood pressure contain several missing values (indicated by zeros, which are unrealistic), so we removed these samples in the pre-processing. According to the ground truth[3], there are in total three causal tiers on the variables: 1. Age, Diabetes_Pedigree_Function, Pregnancies,

[3] https://github.com/cmu-phil/example-causal-datasets/blob/main/real/pima-diabetes/ground.truth/pima-diabetes.knowledge.txt

and Skin Thickness. 2. BMI, Blood Pressure, Glucose, and Insulin. 3. Outcome. It is known that variables in the former tiers can not cause variables in the latter tiers. The results are shown in Figure 6.

MERIT performs very well on the dataset in the sense that only 3 edges out of 12 detected should be forbidden. In addition, it detects many meaning edges including the 4 outgoing edges from the variable Age, and that the Diabetes_Pedigree_Function and the Glucose are the causes of diabetes. Although two outgoing edges from the Outcome variable should be forbidden, the relationships remain reasonable. HCM however, struggles with the dataset, there are in total 6 edges which should be forbidden. The incoming edges to the Age variable are not true. Moreover, the relation between the Outcome variable and the Age variable may be due to some confounders, and the edge between the Diabetes_Pedigree_Function and the Outcome is reversed.

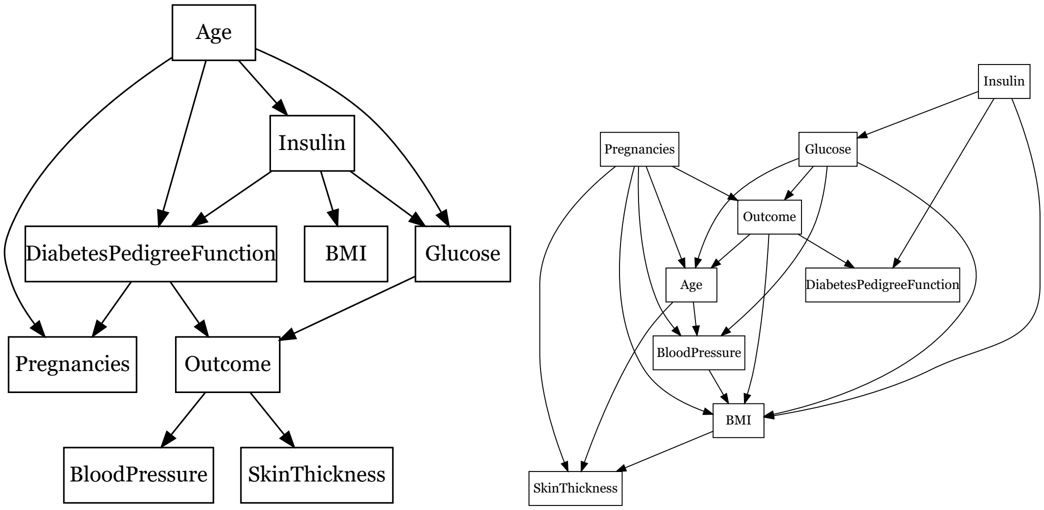

Figure 6: Left: Causal graph estimated by MERIT. Right: Causal graph estimated by HCM.

G.3 Algerian Forest Fire

The dataset was fetched from the UCI repository (https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset), and the ground-truth relationship is provided in https://github.com/cmu-phil/example-causal-datasets/blob/main/real/algerian-forest-fires. We considered all the 12 variables which have known causal tiers, and additionally the target variable Fire for general interest. The known causal tiers are: 1. Region, Month. 2 Temperature, Ws, Rain, RH. 3 FFMC, DMC, DC. 4 ISI, BUI. 5 FWI. Among the variables, Region, Month, and Fire are categorical. A closer examination of the variables reveals that many have long-tailed distributions, including DMC, DC, ISI, BUI,FWI, and rain. Among them, the rain variable is the most severe, with 55% of the data being 0. As this can affect the regression analysis and the independence tests, we conducted experiments on both the original dataset and the transformed dataset where rain is encoded as a bivariate variable, and the other skewed variables are log-transformed (this does not change the ground truth). The results are shown in Figure 7.

In total, MERIT outperformed HCM by having 9 false edges against 11 false edges. More importantly, it correctly identifies most connections with the sink node FWI while HCM regarded it as a parent node to a bunch of variables. Moreover, Region and month are falsely discovered as sink nodes by HCM. We believe the experiments have illustrated the power of the proposed algorithm on different real datasets.

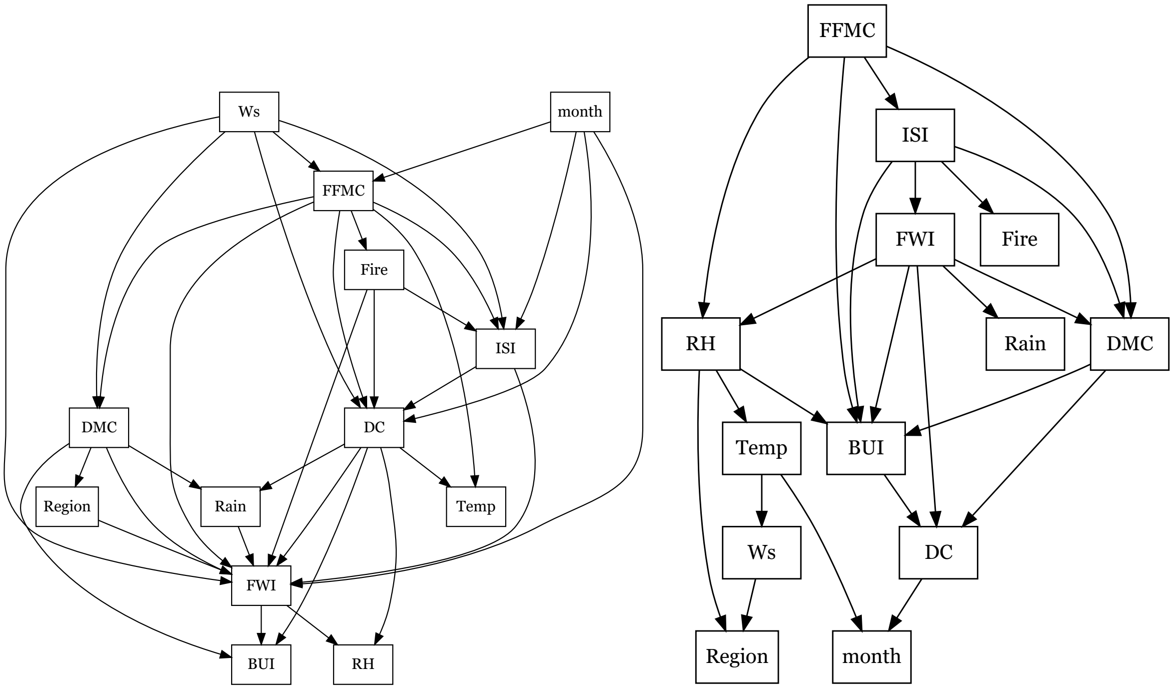

Figure 7: Left: Causal graph estimated by MERIT. Right: Causal graph estimated by HCM.