Nonlinear Causal Structure Learning for Mixed Data

Wenjuan WEI Lu FENG
NEC Labs China
Beijing, China
wei_wenjuan@nec.cn, feng_lu@nec.cn

***Abstract*—Causal discovery from observational data is a fundamental problem. A large number of algorithms have been proposed over the years for that purpose, but they usually handle the data of a single type, either continuous or discrete variables only. Recently, a few causal structure discovery algorithms have been developed for mixed data types, and received many applications. In this paper, we propose a structural equation model for mixed data types, which allows the causal mechanisms to be nonlinear and can consequently model many read-world situations. We prove that the causal structure is identifiable from the data distribution generated by the model under certain conditions. Moreover, we propose a maximum likelihood estimator and develop an efficient order search algorithm benefiting from a novel method of order space cutting, which can handle several hundred variables. We adopt automatic relevance determination kernel-based variable selection after order learning to recover the causal structure. Experiments on synthetic datasets demonstrate the accuracy and scalability of our approach. Especially, we apply our method to publicly available causal-effect pairs and show its superiority in the causal direction identification of mixed causal pairs. In addition, we show that our method can sensibly recover causal relationships on a publicly available real dataset and a private real-world dataset.**

***Index Terms*—causal discovery, structural equation model, observational data, mixed data, nonlinear**

I. Introduction

Causal discovery is well recognized as a challenging yet powerful data analysis tool [1], [2]. The intrinsic appeal of such methods is that they allow us to uncover the causal structure in complex systems, providing an explicit description of the underlying generative mechanisms. Although interventions or randomized experiments supply the golden standard for causal discovery, such approaches are unfeasible or unethical in many scenarios. Alternatively, one may recover causal relationships from passively observational data under proper assumptions.

By assuming the Markov condition and faithfulness, the constraint-based approaches like PC [1] can identify causal skeleton from the joint distribution using conditional independent tests and orient the edges up to the Markov equivalence class via a series of rules (e.g. identifying v-structures, avoiding cycles, etc.). The score-based approaches like GES [3] use a score to evaluate the goodness of fit of candidate causal structures to the data, and output one or multiple graphs with the optimal score. Another line of research is based on structural equation models (SEMs). Such methods [4]–[8] assume that the data generating process belongs to a particular model class describing the causal mechanisms and data distributions, and accept the causal graph fit to the data within the model class.

Most of these methods rely on restrictive assumptions regarding a single data type: continuous [4], [7], [8] or discrete [3], [9], [10]. However, it is always the case to learn the causal structure from a combination of continuous and discrete variables in practice. And several works has been done to relax such assumption about data type [11]. Constraint-based methods have the advantage that they are generally applicable, and thus, in principle, they can be applied to mixed data. The copula PC assumes a combination of continuous and ordinal/binary data are drawn from a Gaussian copula model, and proposes a two-step approach: first estimate a correlation matrix, and then run a causal search algorithm (i.e. PC-Stable) on the estimated correlation matrix [12]. Gibbs sampling is used to estimate the correlation matrix in the first step, and limits the scalability of this method. Mixed graphical models (MGM), originally introduced for learning undirected graphs [13], was adapted for finding directed graphs by adding the post-processing using PC-Stable with likelihood ratio test (LRT) as conditional independence test [14]. Linear regression and logistic regression is used if the dependent variable is continuous or categorical respectively, however linear regression and logistic regression will give different test results for these continuous-discrete edges. Adapting score-based methods to mixed data is a challenging problem [14], and attracts more and more attention. The conditional Gaussian (CG) score partitions the instances of the data according to the values of the discrete variables, and then assumes the continuous variables follow a distinct multivariate Gaussian distribution within each partition [15]. Repeatedly partitioning the data leads to the high computation cost and probably small sample size for some settings of the discrete variables, especially when the number of discrete variables is large. To overcome the drawbacks, the degenerate Gaussian (DG) score was proposed by embedding the discrete variables into a degenerate continuous space and assuming the latent variables in such space to be jointly Gaussian [16]. The MIC score was proposed for mixed data but only allows linear relations [17]. The generalized score functions characterize the (conditional) independence relationships among variables in the general case as a model selection problem in reproducing kernel Hilbert space (RKHS) [18]. It repeatedly conducts kernel ridge regressions and tends to be inefficient for large sample size.

In this paper, we propose a general functional model for mixed-type observed variables, which allows the causal

Corresponding author: Lu FENG.

mechanisms to be nonlinear and hopefully enable much broader real-world applications. For example, the relationship between the dosage and efficacy of a drug has a large chance to be nonlinear and non-monotonic. The proposed model explicitly describes the generating process of a discrete variable as discretization of the combination of a nonlinear effect of its direct causes and an additional distortion. We prove that the causal structure is identifiable from the data distribution following the model under rather mild conditions. We show the identifiability between a mixed pair of continuous and binary variables, and demonstrate how the orientation of a mixed pair may benefit the orientation between a pair of binary variables. Specifically, although the causal direction of two binary variables is not identifiable, we show that introducing an additional continuous variable which has an arrow into one of two binary variables will facilitate the orientation between two binary variables. These theoretical results have been shown to be useful in practical scenarios. A maximum likelihood estimator is developed to learn the causal order among variables instead of the causal structure. If the order is known, the causal structure learning boils down to variable selection [8]. We adopt an automatic relevance determination kernel-based variable selection method to recover the causal structure. To further accelerate the order search, we first learn a sparse skeleton among variables using the graph lasso [19] equipped with kernel alignment, and then project the skeleton onto a series of topological ordering constraints to cut down the search space. Empirical results on synthetic and real data demonstrate the accuracy and scalability of our proposed approach.

II. Mixed Nonlinear Causal Modeling

We start with the definition of a mixed nonlinear causal model (MNCM), and then prove that, if the joint distribution follows such model class, then the causal graph can be identified from the distribution under certain conditions.

A. Model definition

We assume the observed data $\mathbf{X} = (X_1, \cdots, X_D)$ to be a mixture of continuous and binary variables, with no hidden variables. Here we simplify the data types by standard practice of translating a categorical variable with T classes into $(T-1)$ binary variables. We assume the distribution of $\mathbf{X}$ is Markov with respect to an underlying causal DAG $\mathcal{G} = (V, \mathcal{E})$ consisting of nodes $V := \{1, \cdots, D\}$ and edges $\mathcal{E} \subseteq V^2$. Each random variable X_i corresponds to the i-th node in $\mathcal{G}$, and $(i, j) \in \mathcal{E}$ if X_i is a direct cause of X_j. Throughout the paper, we denote the parent set of the i-th node as PA_i. We use the lower-case letter x_i to represent an observation of the random variable X_i. We assume that the observed data was generated by the following model:

Definition 1: **(Mixed Nonlinear Causal Model [MNCM])** *A mixed nonlinear causal model* [1] *is a tuple $(\mathcal{S}, p(\boldsymbol{\epsilon}))$ over the observed data $\mathbf{X}$, where $\mathcal{S} = (S_1, \cdots, S_D)$ is a collection of D equations,*

$$S_i : X_i = \begin{cases} f_i(X_{\mathrm{PA}_i}) + \epsilon_i, & \text{if } X_i \text{ is continuous} \\ \begin{cases} 1, & f_i(X_{\mathrm{PA}_i}) + \epsilon_i > 0 \\ 0, & \text{otherwise} \end{cases}, & \text{if } X_i \text{ is binary} \end{cases} \tag{1}$$

and $p(\boldsymbol{\epsilon}) = p(\epsilon_1, \cdots, \epsilon_D) = \prod_{i=1}^{D} p(\epsilon_i)$ is the joint distribution of noise variables, where $p(\epsilon_i)$ is Gaussian, specifically, $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \sigma_i > 0$ $(i = 1, \cdots, D)$. $f_i(\cdot)$ is a three times differentiable nonlinear function (possibly different for each i), and it is a constant if and only if the parent set is empty, and it is not constant in any of its arguments, i.e., $\forall i$, and $j \in \mathrm{PA}_i$, there are some $x_{\mathrm{PA}_i \backslash j}$ and some $x_j \neq x_j'$ such that $f_i(x_{\mathrm{PA}_i \backslash j}, x_j) \neq f_i(x_{\mathrm{PA}_i \backslash j}, x_j')$. We assume that the corresponding causal graph is acyclic.

Suppose there are N observations for random variables $\mathbf{X}$. Then based on the definition of MNCM, the joint distribution $p(\mathbf{X})$ is,

$$\begin{aligned} p(\mathbf{X}) &= \prod_{i=1}^{D} p_b(X_i | X_{\mathrm{PA}_i})^{z_i} p_c(X_i | X_{\mathrm{PA}_i})^{(1-z_i)} \\ &= \prod_{i=1}^{D} \prod_{n=1}^{N} \left(1 - \Phi\left(\frac{f_i(x_{\mathrm{PA}_i n})}{\sigma_i}\right)\right)^{(1-x_{in}) z_i} \\ &\quad \Phi\left(\frac{f_i\left(x_{\mathrm{PA}_i n}\right)}{\sigma_i}\right)^{x_{in} z_i} \left(\frac{1}{\sigma_i} \varphi\left(\frac{x_{in} - f_i(x_{\mathrm{PA}_i n})}{\sigma_i}\right)\right)^{(1-z_i)}, \end{aligned} \tag{2}$$

where $p_b(\cdot)$ and $p_c(\cdot)$ denotes the probability distribution of binary and continuous variables, respectively. $z_i \in \{0, 1\}$ is an indicator variable that $z_i = 1$ if the variable X_i is binary and $z_i = 0$ otherwise. x_{in} is the n-th observation of X_i, and $x_{\mathrm{PA}_i n}$ is the n-th observation of X_{PA_i}. $\varphi(\cdot)$ is the density of standard normal distribution, and $\Phi(\cdot)$ is cumulative standard normal distribution function.

B. Identifiability

We first illustrate the basic identifiability principle of the MNCM described as (1) on two variables X and Y. For the case that both two variables are continuous, the model degenerates to the nonlinear additive noise model (ANM) of which the identifiability has been well proved by existing work [5]. Here we only consider the case that one is continuous and the other is binary, and the case that both variables are binary.

X is continuous and Y is binary. We observe that the marginal distribution of X in a forward $X \to Y$ model and that in a backward $Y \to X$ model are different, and this difference can be used to identify the causal direction in the mixed causal pair, which is summarized formally in the following theorem.

Theorem 1: Let the joint distribution of a mixed causal pair X(continuous) and Y (binary) be generated by an MNCM in (1)*. Then the causal direction between X and Y is identifiable from the joint distribution.*

Proof: It is obvious that the marginal distribution $p(X)$ in a forward $X \to Y$ model is Gaussian. In a backward model,

[1] The model looks like PNL [6] at the first glance. However, PNL assumes the outer nonlinear transformation to be invertable which is very important for the proof of identifiability. However, as (1) shown, MNCM needs a discretization function to generate a binary variable, and this discretization function is obviously not invertible.

$p(X) = p_\epsilon(X - g(Y=1)p(Y=1)) + p_\epsilon(X - g(Y=0)p(Y=0))$, where $g(\cdot)$ is a non-constant nonlinear function and thus $g(Y=0) \neq g(Y=1)$, then $p(X)$ follows a Gaussian mixture distribution in the backward model. ■

The proof of Theorem 1 relies on the assumption of a specific noise distribution, i.e., Gaussian. However, the following theorem shows that the causal direction of a mixed pair is identifiable for generic choices of noise distribution. Additionally, we empirically demonstrate this finding on the synthetic data with non-Gaussian noises and the publicly available mixed causal pairs, and the results can be found in Section IV.

Theorem 2: Suppose functions $f(\cdot)$ and $g(\cdot)$ are not constant in any of their arguments, p_X is the probability density function of X, and F_ϵ is the cumulative distribution function of ϵ. Let a mixed causal pair X (continuous) and Y (binary) are generated by

$$Y = \begin{cases} 1, & f(X) + \epsilon > 0 \\ 0, & \textit{otherwise} \end{cases}, \tag{3}$$

where X and ϵ are independent. If there is a backward model, i.e., $X = g(Y) + \epsilon'$, where Y and ϵ' are independent, then the triple (f, p_X, F_ϵ) must satisfy the following equation for all x with constant values $c_1 \neq 0, c_2 > 0$:

$$p_X(x)(1 - F_\epsilon(-f(x))) = c_2 p_X(x + c_1) F_\epsilon(-f(x + c_1)). \tag{4}$$

Proof: Here we provide some intuition rather than a formal proof, which can be found in the appendix. The conditional distributions $p(X|Y=1)$ and $p(X|Y=0)$ in the backward model have the same shape and are simply shifted by $g(Y)$. To make this property also hold in the forward model, it can be shown that (4) must hold. ■

Loosely speaking, the statement that the triple (f, p_X, F_ϵ) has to satisfy (4) amounts to saying that in the generic case, it is not possible to have a backward model with the same joint distribution, which means that we can identify the causal direction of a mixed pair even without the assumption of Gaussian noises.

Both X and Y are binary. Although the causal direction of two binary variables is not identifiable, we show that introducing an additional variable Z will facilitate the orientation between two binary variables. Now we give the identifiability condition of the causal direction between two binary variables, which are stated in Theorem 3.

Theorem 3: Let $p(X, Y, Z)$ be generated by an MNCM in (1) *where $\{X, Y\}$ is a binary causal pair. If Z is a parent to one and only one of the variables in $\{X, Y\}$, the causal direction between X and Y is identifiable from $p(X, Y, Z)$.*

Proof: Suppose Z is the parent of X. Then there are two possible scenarios to consider if the condition is satisfied:

S1. Z is not adjacent to Y, i.e., $Y - X \leftarrow Z$.

a. If $p(Y|Z) = p(Y)$, which means Y and Z are independent, then the triple is a V-structure, and orient $Y \rightarrow X$.

b. If $p(Y|X, Z) = p(Y|X)$, which means Y and Z are independent given X, then the triple is not a V-structure, and orient $Y \leftarrow X$.

S2. Z is adjacent to Y, that means Y is the parent of Z, i.e., $Y - X \leftarrow Z$ and $Y \rightarrow Z$. Then, according to the acyclic assumption, orient $Y \rightarrow X$.

Therefore, it is easy to see that the direction between X and Y is identifiable if the condition is satisfied. ■

Since we can identify the causal direction of a mixed causal pair by Theorem 1 and 2, it is possible to identify the causal direction between two binary variables in certain scenarios following orientation propagation rules in Theorem 3. Further considering the identifiability between continuous variables and mixed variables, we know that for the multivariate model, if the condition in Theorem 3 holds for any pair of adjacent binary variables, the whole causal structure is identifiable, which is given in the following corollary.

Corollary 1: Let $p(\mathbf{X})$ be generated by an MNCM in (1) *with the underlying true graph $\mathcal{G}_0$. If any pair of adjacent binary variables X_i and X_j generated satisfying the following condition: there exists a variable which is the parent to one and only one of the variables in $\{X_i, X_j\}$, then $\mathcal{G}_0$ is identifiable from $p(\mathbf{X})$.*

Proof: Assuming there are two MNCMs with different underlying DAGs $\mathcal{G}$ and $\mathcal{G}'$ over $\mathbf{X}$ that satisfy $p(\mathbf{X})$. According to proposition 29 in [7], there are two variables in $\mathbf{X}$ that have an inverse edge in $\mathcal{G}$ and $\mathcal{G}'$. i) For the case that both variables are continuous, the model degenerates to a kind of nonlinear ANM model of which the identifiability has been well proved by existing work [7]; ii) For the case of a mixture of binary and continuous variables, the direction is identifiable based on Theorem 1; iii) For the case that both variables are binary, from Theorem 3, if the condition is satisfied, the direction is identifiable. Thus, the whole causal structure is identifiable. ■

If the condition in Corollary 1 does not hold, we may not be able to determine the causal direction between two adjacent binary variables. Thus, in this case, we cannot derive a fully identified causal graph but an equivalence class. Here, we give a formal definition of equivalence class with MNCM, and empirically show that even if the condition is not satisfied, a compact equivalence class can be found later.

Definition 2: **(MNCM Equivalent Class)** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and $\mathcal{G}' = (\mathbf{V}, \mathbf{E}')$ be two DAGs over the same set of variables $\mathbf{V}$. $\mathcal{G}$ and $\mathcal{G}'$ are called MNCM equivalence, if and only if they satisfy the following properties.*

(1) $\mathcal{G}$ and $\mathcal{G}'$ have the same causal skeleton.

(2) For any pair of adjacent variables $V_i, V_j \in \mathbf{V}$, if any of them is continuous, the causal direction between V_i and V_j is the same in $\mathcal{G}$ and $\mathcal{G}'$.

(3) For any pair of adjacent binary variables $V_i, V_j \in \mathbf{V}$, if it satisfies the condition given in Theorem 3, then the causal direction between V_i and V_j is the same in $\mathcal{G}$ and $\mathcal{G}'$.

III. Causal Inference

Although we have proved the true causal structure is identifiable from the joint distribution following an MNCM under the condition in Corollary 1, it is still time-consuming work searching the entire DAG space to find the optimal causal graph. Previous research [20] shows the structure learning

problem with DAG constraint can be cast as that of learning an optimal order among variables, which intuitively looks easier since the order space is much smaller than the DAG space. Once the order was determined, the constraint of no cycle can be enforced by constraining the parents of a variable to be a subset of variables ordering precede it, and the causal structure learning boils down to variable selection.

Here we propose an efficient causal discovery approach with two stages: the causal order learning and the causal structure recovery. A preliminary order-space cutting is recommended to gain a higher efficiency.

A. Causal Order Learning

We begin with order identifiability, and then find the optimal order using the maximum likelihood estimation.

1) Order identifiability: We denote all the permutations on $\{1, \cdots, D\}$ as $\mathcal{P}_D$, and for each permutation $\xi \in \mathcal{P}_D$, we define the fully connected DAG $\mathcal{G}_\xi^{full}$ as the DAG containing all edges $i \to j$ for $\xi(i) < \xi(j)$. The following corollary gives the order identifiability formally.

Corollary 2: Let $p(\mathbf{X})$ be generated by an MNCM in (1) *with the underlying true graph $\mathcal{G}_0$. Without the assumption that f is not constant in any of its arguments, and if the condition in Corollary 1 holds, the set of true orderings, $\Xi^0 := \{\xi \in \mathcal{P}_D | \mathcal{G}_\xi^{full} \geq \mathcal{G}_0\}$ is identifiable from $p(\mathbf{X})$, where $\mathcal{G}_\xi^{full}$ is the supergraph of $\mathcal{G}_0$.*

Proof: The proof is similar to the proof of Lemma 32 in [7], and thus we only provide an intuition here. Without the assumption that f is not constant in any of its arguments, we can set $f_j(X_{\mathrm{PA}_j \setminus i}, X_i) = f_j(X_{\mathrm{PA}_j \setminus i})$ if X_i has no effect on X_j. Then for a fully connected DAG $\mathcal{G}_\xi^{full}$ leading to $p(\mathbf{X})$, we can find the minimal graph $\mathcal{G}_\xi^{full,min}$ that leads to $p(\mathbf{X})$ and satisfies the condition that f is not constant in any of its arguments. If the condition in Corollary 1 holds, $\mathcal{G}_0$ is identifiable from $p(\mathbf{X})$, then $\mathcal{G}_\xi^{full,min} = \mathcal{G}_0$. ■

2) Order Estimation: The log-likelihood of MNCM on N observations of $\mathbf{X}$ and an order ξ is,

$$\mathcal{L}_{\mathrm{MNCM}}(\mathbf{X}, \mathcal{G}_\xi^{full}) = \sum_{i=1}^{D} \sum_{n=1}^{N} z_i \left(x_{in} \log \left(\Phi \left(\frac{f_i(x_{\mathrm{PA}_i,n})}{\sigma_i} \right) \right) + (1 - x_{in}) \log \left(1 - \Phi \left(\frac{f_i(x_{\mathrm{PA}_i,n})}{\sigma_i} \right) \right) \right) - \sum_{i=1}^{D} \frac{(1 - z_i)N}{2} \left(\log \frac{\sum_{n=1}^{N} (x_{in} - f_i(x_{\mathrm{PA}_i,n}))^2}{N} \right) + \mathrm{const}, \tag{5}$$

where PA_i is the parent set of i-th node in the fully connected DAG $\mathcal{G}_\xi^{full}$ based on order ξ.

We use a greedy search to learn the optimal order $\hat{\xi}$ which maximizes the unpenalized log-likelihood,

$$\hat{\xi} = \arg\max_{\xi} \mathcal{L}_{\mathrm{MNCM}}(\mathbf{X}, \mathcal{G}_\xi^{full}). \tag{6}$$

The search procedure is similar to CAM [8]. Start with an empty DAG and at each iteration add one edge $i \to j$ corresponding to the largest gain of $\mathcal{L}_{\mathrm{MNCM}}$. Then check the acyclicity after each iteration with the order information of $i \prec j$, and after the addition of an edge, we only need to recompute the score related to j, and construct a super DAG after all the iterations.

B. Causal Structure Recovery

We apply the automatic relevance determination (ARD) kernel-based Gaussian process regression [21] or classification [22] for each variable to prune the super DAG. Specifically, for a variable X_i with its parents PA_i in the super DAG, the ARD kernel used for X_i is

$$\mathbf{K}_{\mathrm{ARD}} = \prod_{j \in \mathrm{PA}_i} \mathbf{K}_j, \tag{7}$$

where $\mathbf{K}_j$ is the centralized kernel matrix of X_j whose (n, n')-th element is $\mathbf{K}_j(n, n') = \exp\left(-\frac{||x_{jn} - x_{jn'}||^2}{2l_j^2}\right)$.

With an ARD kernel, each feature comes with an independent length scale l_j. When optimizing l_j, some of them could be concentrated at large value along with their corresponding features eliminated, resulting that highly relevant features can be effectively extracted.

C. Order Search Space Cutting

We utilize the ancestor and neighbour relations to further cut down the order search space to accelerate the estimation procedure. Obviously, the ancestor relations can greatly prune the search space. Suppose we already know an ancestor relation, i.e., i precedes j. Then we do not have to consider those orders in which j precedes i. As for the neighbour relations, the following theorem gives the efficacy of them in the search space cutting.

Theorem 4: Suppose that starting from an initial potential parents of i which are set to $V \setminus \{i\}$, $i \in \{1, \cdots, D\}$, a fully connected DAG with the true order ξ, i.e., $\mathcal{G}_\xi^{full}$, can be recovered. If the initial potential parents of i were set to NB_i, the neighbours of i, then a DAG $\mathcal{G}_\xi^1$ can be recovered with $\mathcal{L}_{MNCM}(\mathbf{X}, \mathcal{G}_\xi^{full}) = \mathcal{L}_{MNCM}(\mathbf{X}, \mathcal{G}_\xi^1)$.

Proof: Let PA_i denote the true parent set of i. Then $\mathrm{PA}_i = \mathrm{PA}_i^{\mathcal{G}_\xi^1} \subseteq \mathrm{PA}_i^{\mathcal{G}_\xi^{full}}$. Let $S := \mathrm{PA}_i^{\mathcal{G}_\xi^{full}} \setminus \mathrm{PA}_i^{\mathcal{G}_\xi^1}$. Then we have $X_i \perp\!\!\!\perp X_S | X_{\mathrm{PA}_i}$ based on the Markov condition. Therefore, $\log p(X_i | X_{\mathrm{PA}_i}) = \log p(X_i | X_{\mathrm{PA}_i}, X_S)$, and then $\mathcal{L}_{\mathrm{MNCM}}(\mathbf{X}, \mathcal{G}_\xi^{full}) = \mathcal{L}_{\mathrm{MNCM}}(\mathbf{X}, \mathcal{G}_\xi^1)$. ■

There are many ways to learn the ancestor or neighbour relations, currently we adopt a method based on kernel alignment and graph lasso. The kernel alignment was originally applied to measure the similarity between two kernel functions [23]. And recently it was used to produce the pseudo-correlation matrix among random variables [24]. We generate the pseudo-correlation matrix $\mathbf{A}$ over the observed data $\mathbf{X} = (X_1, \cdots, X_D)$ using (8) with each element $\mathbf{A}(i, j)$ being the kernel alignment between X_i and X_j:

$$\mathbf{A}(i, j) = \frac{\langle \mathbf{K}_i, \mathbf{K}_j \rangle}{\sqrt{\langle \mathbf{K}_i, \mathbf{K}_i \rangle \langle \mathbf{K}_j, \mathbf{K}_j \rangle}}, \tag{8}$$

where $\langle \mathbf{K}_i, \mathbf{K}_j \rangle = \sum_{n,n'=1}^{N} \mathbf{K}_i(n, n') \mathbf{K}_j(n, n')$, and $\mathbf{K}_i(n, n')$ is the (n, n')-th element of the centralized kernel

matrix of X_i. Here we use the radial basis function (RBF) kernel for continuous variables and delta kernel for binary variables. Then introduce $\mathbf{A}$ into the graph lasso to learn the precision matrix $\mathbf{\Theta}$ as

$$\mathbf{\Theta} = \arg\min_{\mathbf{\Theta}\succ 0} \operatorname{tr}(\mathbf{A}\Theta) - \log\det(\mathbf{\Theta}) + \lambda \sum_{i,j} |\mathbf{\Theta}_{ij}|. \quad (9)$$

$\mathbf{\Theta}_{ij} = 0$ means there is no direct edge between X_i and X_j.

The strongly connected components (SCCs) [25] are generated from $\mathbf{\Theta}$ with no edge connecting different SCCs, and thus the topological orders among SCCs can be arbitrarily assigned. If $SCC_m \prec SCC_{m'}$, then $X_i \prec X_j$ for all $X_i \in SCC_m$ and $X_j \in SCC_{m'}$. These order constraints can be used to cut down the search space.

D. Algorithm and Implementation

The whole algorithm is summarized in Algorithm 1. First we learn a sparse skeleton among variables using the graph lasso equipped with the kernel alignment method, and then project this skeleton onto a series of topological ordering constraints to cut down the search space. Next, estimate $\hat{\xi}$ in (6) using a greedy search over the feasible space. We utilize the Gaussian process regression (classification) to estimate $\log p(X_i|X_{\mathrm{PA}_i})$ during the search. Finally, recover $\mathcal{G}_{\hat{\xi}}^{full,min}$ from $\mathcal{G}_{\hat{\xi}}^{full}$ by pruning edges with the ARD kernel-based Gaussian process regression (classification) for each variable, and the parents are chosen as those variables whose length scale is smaller than a predefined threshold. A standard gradient descent optimizer is used to fit the hyperparameters of Gaussian processing regression (classification) through maximizing the log marginal likelihood. The time complexity of our order search algorithm is $O(M \max_m |SCC_m| N^3)$, where M is the number of SCCs, $|SCC_m|$ is the number of edges in SCC_m according to $\mathbf{\Theta}$, $m \in \{1, \cdots, M\}$, and N is the sample size.

Algorithm 1 $\mathcal{L}_{\mathrm{MNCM}}$-based causal structure discovery

Input: Data $\boldsymbol{X}$, the number of variables D, threshold α
Output: Optimal structure $\hat{\mathcal{G}} \in \{0,1\}^{D\times D}$, causal order $\hat{\xi}$

PHASE 1: Order search space cutting
Construct the precision matrix $\mathbf{\Theta}$ using (8) and (9).
Extract M SCCs using Tarjan's algorithm from $\mathbf{\Theta}$.
Assign an arbitrary group order e.g. $SCC_1 \prec \cdots \prec SCC_M$ and construct order constraints set $\mathbf{C}$.
PHASE 2: Causal order learning
Initial $t = 1$, an empty DAG $\hat{\mathcal{G}} = \mathbf{0}$, a score matrix $\boldsymbol{S} = \{-\inf\}^{D\times D}$, and a node score list $\mathbf{NS} = \{-\inf\}^{1\times D}$.
Compute $\mathbf{NS}^t[j] = \log p(X_j|\emptyset)$.
Compute $\boldsymbol{S}^t[i,j] = \log p(X_j|X_i) - \mathbf{NS}^t[j]$, if $\mathbf{\Theta}_{ij} \neq 0$.
if $i \prec j$ violates $\mathbf{C}$ **then** $\boldsymbol{S}^t[i,j] = -\inf$.
for $m = 1, \cdots, M$ **do**
  while *TRUE* **do**
    Find $(\hat{i}, \hat{j}) = \arg\max_{i,j\in SCC_m} \boldsymbol{S}^t[i,j]$.
    if $\boldsymbol{S}^t[\hat{i},\hat{j}] = -\inf$ **then** break.
    Set $\mathcal{G}[\hat{i},\hat{j}] = 1$, and add $\hat{i} \prec \hat{j}$ to $\hat{\xi}$.
    Set $\boldsymbol{S}^t[i,j] = -\inf$, $\forall i,j \in SCC_m$ that violate acycle.
    $\mathbf{NS}^t[\hat{j}] = \boldsymbol{S}^t[\hat{i},\hat{j}] + \mathbf{NS}^t[\hat{j}]$.
    $t = t + 1$.
    Update $\boldsymbol{S}^t[i,\hat{j}] = \log p(X_{\hat{j}}|X_{\mathrm{PA}_{\hat{j}}}, X_i) - \mathbf{NS}^{t-1}[\hat{j}]$ if $\boldsymbol{S}^{t-1}[i,\hat{j}] \neq -\inf\ \forall i \in SCC_m$.
  end while
end for
PHASE 3: Causal structure recovery
for $j = \{1, \cdots, D\}$ **do**
  $\mathrm{PA}_j = \{i \mid \hat{\mathcal{G}}[i,j] = 1\}$.
  Learn $\hat{l}_{\mathrm{PA}_j}$ from (7) based on X_j and X_{PA_j}.
  Find the non-relevant features as $\mathrm{NPA}_j = \{\hat{l}_{\mathrm{PA}_j} \geq \alpha\}$.
  Set $\hat{\mathcal{G}}[\mathrm{NPA}_j, j] = 0$.
end for

IV. Experiments

We conducted various experiments to have a clear understanding of the MNCM and our method. First, we used some toy examples to empirically verify Corollary 1. Then, we compared our method with several benchmarks on various causal structures when the condition in Corollary 1 holds or not. Thirdly, we evaluated the robustness of our method on non-Gaussian distribution noises. We applied our method to publicly available mixed causal pairs to empirically verify Theorem 2, and also applied it to a publicly available real dataset and a private real-world dataset.

A. Simulated Study

1) Synthetic data: **Dataset 1** was used to empirically verify Corollary 1 and to illustrate how a mixed pair benefits the orientation between a binary pair. We generated three variables X, Y, Z with two different structures, in which X and Y are binary and Z is continuous.

Case A: $Z \to X \to Y$ when the condition in Corollary 1 holds.

Case B: $Z \to X \to Y$ and $Z \to Y$ when the condition in Corollary 1 does not hold.

Z was generated following a Gaussian distribution with a standard deviation uniformly sampled from $[1, \sqrt{2}]$. X and Y were generated by an MNCM in (1), in which f_i was a Gaussian process with a bandwidth one RBF kernel, ϵ_i was a Gaussian noise with a standard deviation uniformly sampled from $[\frac{1}{5}, \frac{\sqrt{2}}{5}]$, and we chose the cutting value based on p-th percentile of the continuous value, where p was randomly chosen from $[10, 90]$. We generated 50 datasets for each case and the sample size of each dataset was 500.

Dataset 2 was generated when the condition in Corollary 1 held. First we randomly generated causal structures, i.e. DAGs with $D = 10, 30, 50, 100$ and different graph densities, which are measured by the expected number of edges. The sparse graphs have D edges while the dense graphs own $2D$ edges on average. Note that those causal structures violating the condition were eliminated. After that, the data were generated from an MNCM with $N = 500, 1000$ and different binary ratios $0.1, 0.5$, which are measured by the ratio of binary variables in total variables. We drawn the functions f_i from a Gaussian process with a bandwidth one RBF kernel, and added Gaussian noise ϵ_i with a standard deviation uniformly sampled from $[1, \sqrt{2}]$ for the nodes without parents, and uniformly sampled from $[\frac{1}{5}, \frac{\sqrt{2}}{5}]$ for the other nodes, following [8]. As for the

binary variable, we chose the cutting value based on p-th percentile of the continuous value, where p-th was randomly chosen from $[10, 90]$.

Dataset 3 was generated when the condition in Corollary 1 did not hold. The data generating process was the same as Dataset 2 except that the condition was violated in each causal structure. We only set $D = 10, 30$, $N = 1000$, and the binary ratio was 0.5, due to the time limit.

Dataset 4 was generated in the same way as dataset 2, except the noise distributions were non-Gaussian. Specifically, for any pair of adjacent binary pair in the graph, there was a variable which was the parent to one and only one of the binary pair. Here, we considered two types of noise distributions.

(1) *Uniform distribution*: ϵ_i was uniformly sampled from $[-a, a]$, where a was uniformly sampled from $[3, 4]$ for the nodes without parents and from $[0.5, 1]$ for the other nodes.

(2) *Gaussian mixture distribution*: $\epsilon_i = p_1\mathcal{N}(\mu, \sigma^2) + (1 - p_1)\mathcal{N}(-\mu, \sigma^2)$ with $p_1 = 0.5$, μ was uniformly sampled from $[1, 2]$ and σ was uniformly sampled from $[1, \sqrt{2}]$ for the nodes without parents, while μ was uniformly sampled from $[0.1, 0.5]$ and σ was uniformly sampled from $[\frac{1}{5}, \frac{\sqrt{2}}{5}]$ for the other nodes.

2) Benchmarks: On the dataset 2, 3, and 4, we compared our method with 5 benchmarks, including the score-based ones: 1) GES using conditional Gaussian BIC score (GES_CGBIC) [15], 2) GES using degenerated Gaussian score (GES_DGBIC) [16], 3) GES using the generalized score (GES_GS) [18], and the constraint-based ones: 4) copula PC [12], 5) causalMGM [14]. We used the implementations provided by authors and their default parameter settings for the benchmarks. For the constraint-based methods, we used the PC-Stable with alpha levels of 0.05. For our method, on both synthetic data and real data, we used the kernel width twice the median distance between points in the first and second phases of Algorithm 1, and used the BIC score to select the λ and the related precision matrix $\mathbf{\Theta}$. As for the pruning phase, the hyperparameter α was selected from $\{1, 2, 3\}$ by cross validation.

3) Evaluation metrics: For dataset 1, we measured the accuracy as how many times we recovered the true causal structure or true MNCM equivalence class out of 50 random datasets.

For dataset 2 and dataset 4, we measured the accuracy of causal structure discovery with three metrics: 1) the $F1$ score, which is a weighted average of the precision and recall, 2) the normalized structural hamming distance (SHD) [26], which counts how many edges whose types do not coincide between two graphs, 3) the structure intervention distance (SID), which is well suited for quantifying the correctness of an order among variables [27]. The SID weights a reversed edge in the estimated DAGs greater than an additional edge, while SHD weights both errors equally. Considering our method outputs DAGs, while the benchmarks output CPDAGs, we selected the best DAG within the Markov equivalent class to represent the performance of benchmarks.

For dataset 3, since our method ouputs the MNCM equivalence class while the benchmarks output CPDAGs, we provided the upper bound and lower bound of $F1$ score, SHD, and SID within equivalence classes.

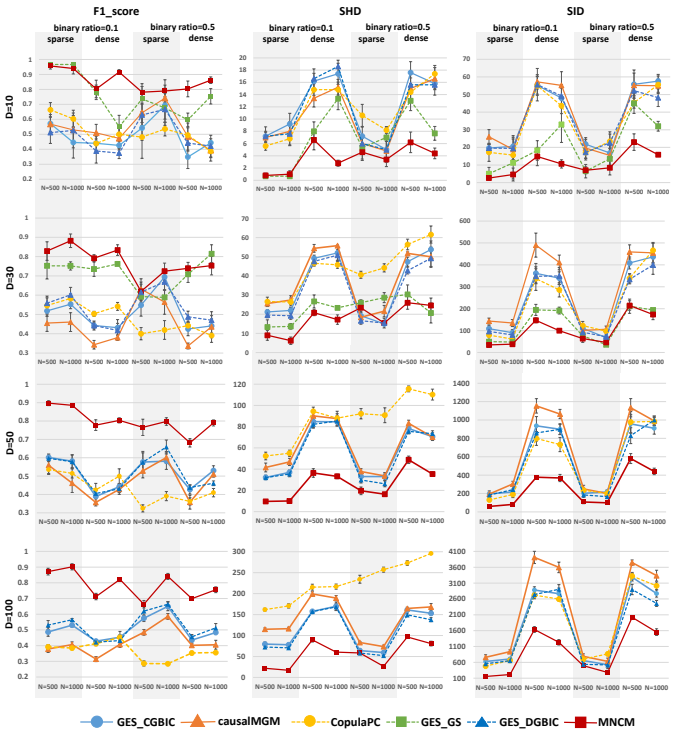

Fig. 1. Comparisons of causal structure discovery on **Dataset 2**. The $F1$ score (left column), SHD (middle column) and SID (right column) of different methods were illustrated on different settings varying the numbers of nodes (different rows of figure), binary ratios, graph densities, and sample sizes (x-axis of each subgraph).

4) Results: **Results on dataset 1:** In the case A, our method recovered the true causal structure at most of the time (the accuracy was 0.88), and reversed the edge between X and Y, i.e. $Z \to X \leftarrow Y$, at rest of the time. In the case B, our method recovered the true causal structure with a probability of 60%, and found the true MNCM equivalence class with a probability of 100%. Our results verified the true causal structure could be recovered under the condition in Corollary 1 with a high probability, and even if the condition didn't hold, at least an MNCM equivalence class could be found. Moreover, the results in case A and case B both verified that our method could identify the direction between mixed pairs.

Results on dataset 2: Our method was compared with benchmarks on different settings when the condition in Corollary 1 held, and the results were summarized in Figure 1. Higher $F1$ scores and lower SHD scores mean higher accuracy on the causal structure discovery, and lower SID values indicate a better performance on the order recovery. As we used the best DAG for the benchmarks, we could consider that we compared the proposed method with the best performance of benchmarks. Overall, the proposed method outperformed others in terms of causal structure discovery. One explanation is that most of the score-based methods have to make assumptions about causal mechanisms and data distributions, which may not hold, and

TABLE I
SIDs OF THE FULLY-CONNECTED DAG. STANDARD DEVIATIONS ARE SHOWED IN PARENTHESES.

SETTINGS	BINARY RATIO=0.1				BINARY RATIO=0.5			
	SPARSE		DENSE		SPARSE		DENSE	
	N=500	N=1000	N=500	N=1000	N=500	N=1000	N=500	N=1000
D=10	0(0)	0(0)	0(0)	2.75(3.20)	0.8(1.10)	0.8(1.15)	3.67(3.50)	3.71(4.11)
D=30	11(8.37)	8.8(9.83)	20.8(13.3)	7(14.04)	17.8(19.00)	6.8(4.09)	18.5(7.72)	16.25(26.84)

the constraint-based methods may suffer from the statistical test issues. Specifically, GES_CGBIC and GES_DGBIC assume the linear relations, causalMGM uses the likelihood ratio test based on linear/logistic regression, and copulaPC supposes a monotonic relationship, and thus these methods performed worse when the underlying causal mechanisms were nonlinear. As GES_GS has no explicit assumption about the relationship and data distribution, it worked better than other benchmarks in general. However, GES_GS did not work on highly dimensional data (e.g. $D = 50$ or 100) due to its high computational complexity. Moreover, the proposed method achieved the lowest SID in most cases, especially for dense graphs and high binary ratio, implying that it can recover the true order. To further demonstrate the superiority of our method in terms of order recovery, we provided the SIDs of the fully-connected DAG resulting from the second phase of Algorithm 1 in Table I. The SID ranges from 0 (which means a correct order) to $D(D-1)$ (which means a totally wrong order). As Table I shows, the proposed method always obtains low SID values.

Results on dataset 3: We evaluated the performance of our method when the condition in Corollary 1 did not hold, and the results were summarized in Fig. 2. Overall, the proposed method outperformed others with a tighter interval between the lower bound and upper bound. Taking a deep look into the data, we found that even though the condition in Corollary 1 didn't hold in every dataset, the proportion of the non-identifiable pairs was not so large. On average, there were about 20 percentage of the binary pairs not satisfying the condition.

Results on dataset 4: We compared our method with benchmarks on different settings and the results were summarized in Fig. 3 for uniform noises and Fig. 4 for Gaussian mixture noises. As we used the best DAG for the benchmarks, we could consider that we compared the proposed method with the upper bound of benchmarks. Overall, the proposed method outperformed others in terms of causal structure discovery even with non-Gaussian noises. One explanation is that GES_CGBIC, GES_DGBIC, and causalMGM assume the relations among variables to be linear, and copulaPC supposes a monotonic relationship, thus these methods performed worse when the underlying causal mechanisms were nonlinear. As GES_GS has no explicit assumption about the relationship and the data distribution, it worked better than other benchmarks in general. However, GES_GS could not work on highly dimensional data (e.g. $D = 50, 100$) due to its high computational complexity. It also implied that the direction of a mixed causal pair could be identified robustly even when the noise distributions were non-Gaussian, and the orientation of a mixed pair might benefit the orientation between a pair of binary variables in certain scenarios.

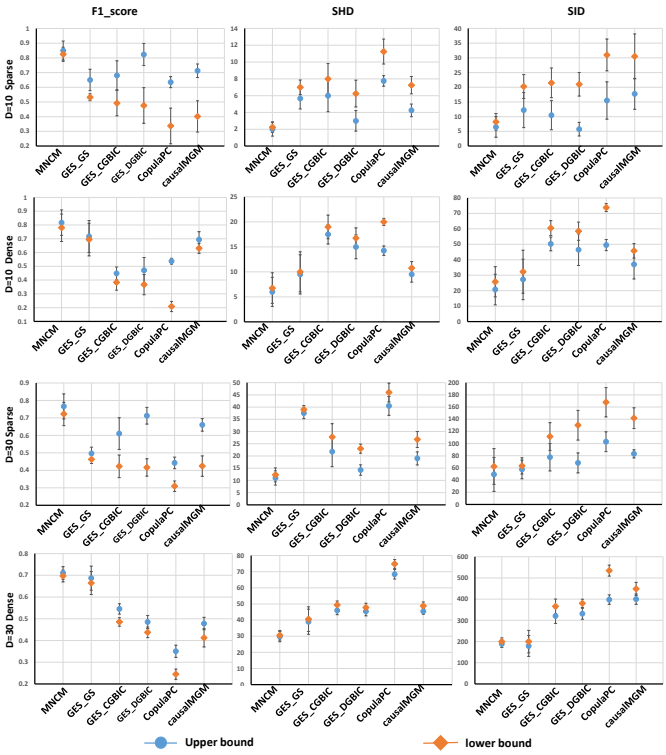

Fig. 2. Comparisons of causal structure discovery on **Dataset 3**. The upper bounds and lower bounds of $F1$ score (left column), SHD (middle column) and SID (right column) between true DAG and the estimated graph of different methods were illustrated on different settings varying the numbers of nodes and graph densities (different rows of figure).

B. Results on Continuous-Binary Pair

We used the causal effect challenge dataset [2] to verify the performance of our method in terms of identifying the direction between continuous-binary pairs. Only the train data were used, and totally there were 295 pairs including 154 pairs with causal directions being from continuous to binary and 141 pairs with causal directions reversing.

We compared our method with 3 other methods, 1) Information Geometric Causal Inference (IGCI) [28], 2) bivariate ANM with GP regression, using HSIC score (ANM-HSIC)

[2]https://www.kaggle.com/c/cause-effect-pairs/data

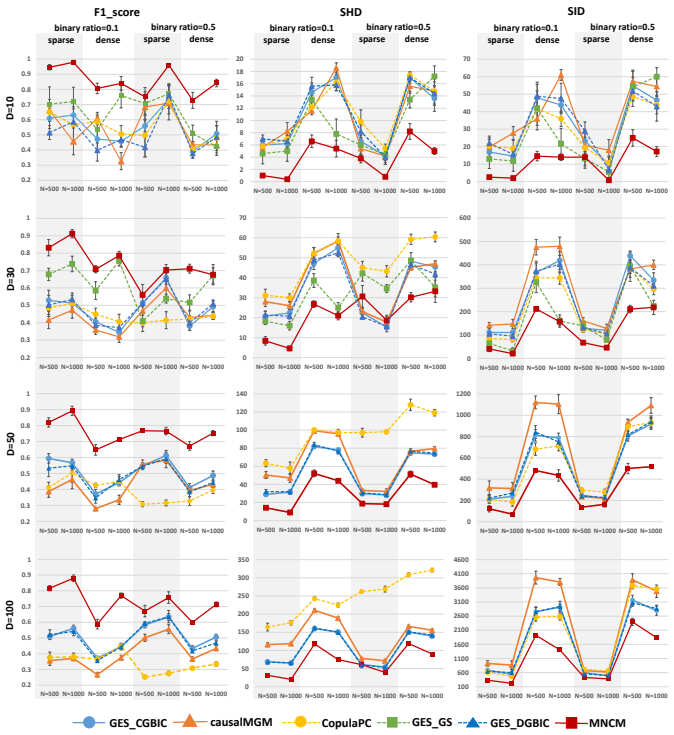

Fig. 3. Comparisons of causal structure discovery on **Dataset 4** with uniform distribution noises. The $F1$ score (left column), SHD (middle column) and SID (right column) of different methods were illustrated on different settings varying the numbers of nodes (different rows of figure), binary ratios, graph densities, and sample sizes (x-axis of each subgraph).

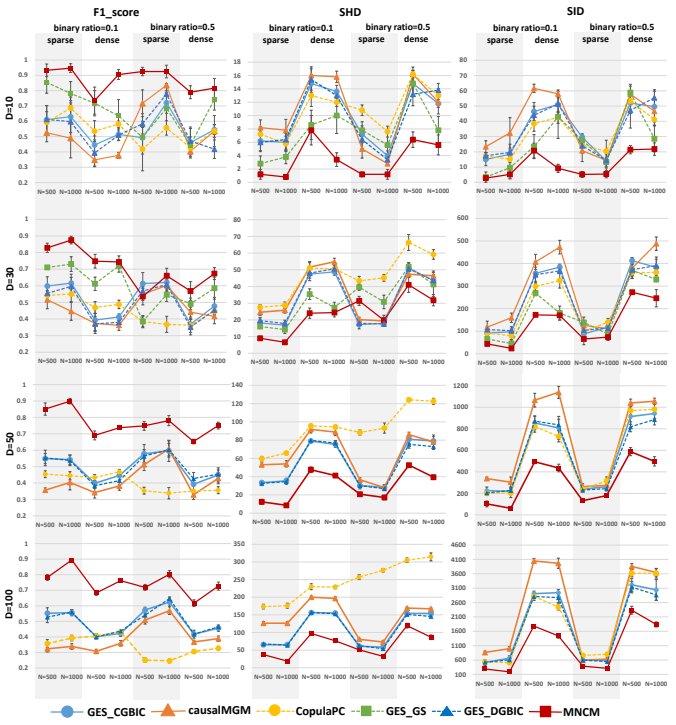

Fig. 4. Comparisons of causal structure discovery on **Dataset 4** with Gaussian mixture distribution noises. The $F1$ score (left column), SHD (middle column) and SID (right column) of different methods were illustrated on different settings varying the numbers of nodes (different rows of figure), binary ratios, graph densities, and sample sizes (x-axis of each subgraph).

[29], 3) bivariate ANM with GP regression, using Gaussian score (Bi-CAM) [8]. All the benchmarks assume both variables to be continuous, and we failed to find a bivariate orientation method for mixed causal pairs. We used the implementations provided by authors and their default parameter settings for the benchmarks.

We measured the accuracy as how many times we recovered the true causal direction out of 295 pairs. Overall, the proposed method outperformed the others in causal direction identification of the continuous-binary pairs (accuracy: MNCM: 0.66, IGCI: 0.456, AMN-HSIC: 0.479 and Bi-CAM: 0.54). One explanation is that all the benchmarks assume both variables to be continuous, that is violated in the mixed pair dataset.

C. Real World Retail Data

We applied our method to a real retail dataset including the store information, sales data ranging from 2017 to 2019, and product attributes of toothpaste products under a brand. The aim was to find the causes driving the success of new products. We found 389 new products launched after 2017, along with 21 features including the product characteristics (i.e., tube or not (binary), benefit (3, binary), has gift or not (binary), regular or not (binary), gel or not (binary)), marketing strategies (i.e. price of the new product (continuous), seasons that the new product launched (3, binary), distributed strength (continuous), store size in different areas when the new product launched (4, continuous), store coverage of new product in different areas (4, continuous)), and whether the new product is success or not (binary).

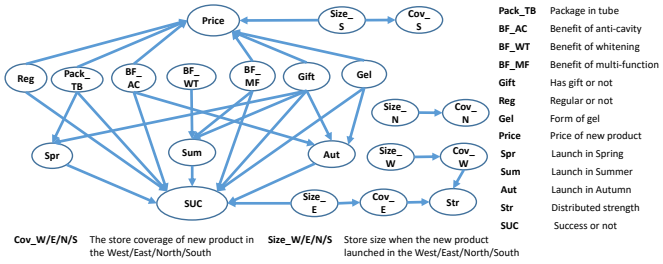

Fig. 5. Recovered causal graph from real world retail data.

Fig. 5 shows the causal graph recovered by our algorithm. It reveals some relations that are in accord with common sense about the market, e.g. new products' characteristics and launch seasons have effects on the success of them, and launch seasons are influenced by some characteristics of new products (e.g. whitening function leads to the summer launch). More interestingly, we find that the price has no effect on the success of new products, but they are correlated due to common parents: the characteristics of new product. Moreover, the store size in the East has an effect on the the success of new product,

which gives us a new idea about distributing strategies when new products launch.

D. Results on Boston Housing Data

We applied our method to Boston housing data[3] which contains 14 variables including 1 binary variable and 506 samples. This dataset has been used in previous studies to evaluate the performance of causal structure discovery in the real application [30], [31]. Fig. 6 shows the recovered causal graph by our algorithm. It confirmed that the number of rooms (RM), percentage of lower-status population (LST), proportion of owner-occupied units built before 1940 (AGE), and crime rate (CRI) are direct causes of the median value of housing price (MED), which were also recovered by [31]. Besides that, we found some other interesting relations which are consistent with our common understandings, e.g. a link from the tax rate (TAX) to the pupil-teacher rate (PTR), the distance to employment centers (DIS) to the index of accessibility to radial highways (RAD).

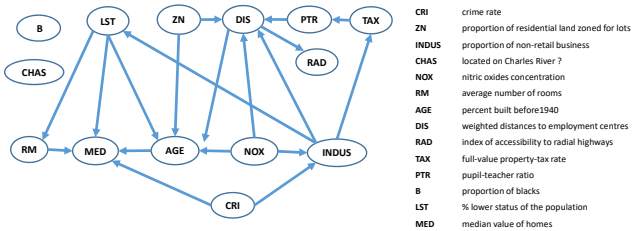

Fig. 6. Recovered causal graph from Boston housing data.

V. Conclusion

In this paper, we proposed a mixed nonlinear causal model to describe the nonlinear relationships among a mixture of discrete and continuous variables, and proved its identifiability under certain condition. We then proposed a maximum likelihood estimator to learn the causal order and with such estimator, we also developed an efficient order search algorithm benefiting from a novel method of order space cutting. Finally, we adopt an automatic relevance determination kernel-based variable selection after order learnt to recovery the causal structure. Empirical results on synthetic and real-world datasets demonstrate the accuracy and scalability of our proposed approach.

Appendix

Proof of Theorem 2: The conditional distributions in a forward $X \to Y$ model are,

$$\begin{aligned} P(X|Y=1) &= \frac{p_X(X)(1-F_\epsilon(-f(X)))}{p(Y=1)}, \\ P(X|Y=0) &= \frac{p_X(X)F_\epsilon(-f(X))}{p(Y=0)}, \end{aligned} \tag{10}$$

where p_X is the probability density function of X, and F_ϵ is the cumulative distribution function of ϵ.

In contrast, the conditional distributions in a backward $Y \to X$ model are,

$$\begin{aligned} P(X|Y=1) &= p_{\epsilon'}(X - g(Y=1)), \\ P(X|Y=0) &= p_{\epsilon'}(X - g(Y=0)), \end{aligned} \tag{11}$$

where $p_{\epsilon'}$ is the probability density function of ϵ'.

In the backward model, the conditional distributions of $p(X|Y=1)$ and $p(X|Y=0)$ have the same distribution and are simply shifted by the function $g(Y)$. If the forward and backward model coexist sharing the same joint distribution, then the forward model must satisfy the following:

for all x, there exists a constant $c_1 = g(Y=0) - g(Y=1) \neq 0$, s.t. $p(X=x|Y=1) = p(X=x+c_1|Y=0)$. That is,

$$\frac{p_X(x)(1-F_\epsilon(-f(x)))}{p(Y=1)} = \frac{p_X(x+c_1)F_\epsilon(-f(x+c_1))}{p(Y=0)}, \tag{12}$$

where $c_1 \neq 0$. Then,

$$\frac{p_X(x)(1-F_\epsilon(-f(x)))}{p_X(x+c_1)F_\epsilon(-f(x+c_1))} = \frac{p(Y=1)}{p(Y=0)} = c_2, \tag{13}$$

where $c_2 := \frac{p(Y=1)}{p(Y=0)} > 0$ is a constant. ■

[3]http://lib.stat.cmu.edu/datasets/boston

References

[1] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT Press, 2000.

[2] J. Pearl, *Causality: models, reasoning and inference, 2nd ed.* Cambridge University Press, 2009.

[3] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 507–554, 2003.

[4] S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. J. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.

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

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

[7] J. Peters, J. M. Mooij, D. Janzing, and B. Scholkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2009–2053, 2014.

[8] P. Buhlmann, J. Peters, and J. Ernest, "Cam: Causal additive models, high-dimensional order search and penalized regression," *The Annals of Statistics*, vol. 42, no. 6, pp. 2526–2556, 2014.

[9] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on discrete data using additive noise models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[10] R. Cai, J. Qiao, K. Zhang, Z. Zhang, and Z. Hao, "Causal discovery from discrete data using hidden compact representation," *Neural Information Processing Systems*, pp. 2666–2674, 2018.

[11] V. K. Raghu, A. Poon, and P. V. Benos, "Evaluation of causal structure learning methods on mixed data types," *Journal of Machine Learning Research*, 2018.

[12] R. Cui, P. Groot, and T. Heskes, "Copula pc algorithm for causal discovery from mixed data," in *European Conference on Machine Learning*. Springer, 2016, pp. 377–392.

[13] J. D. Lee and T. J. Hastie, "Learning the structure of mixed graphical models," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 230–253, 2015.

[14] A. J. Sedgewick, J. Ramsey, P. Spirtes, C. Glymour, and P. V. Benos, "Mixed graphical models for causal analysis of multi-modal variables," *arXiv: Artificial Intelligence*, 2017.

[15] B. Andrews, J. Ramsey, and G. F. Cooper, "Scoring bayesian networks of mixed variables," *International Journal of Data Science and Analytics*, vol. 6, no. 1, pp. 3–18, 2018.

[16] Andrews, Bryan and Ramsey, Joseph and Cooper, Gregory F, "Learning high-dimensional directed acyclic graphs with mixed data-types," *Proceedings of Machine Learning Research*, vol. 104, no. 1, pp. 4–21, 2019.

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

[18] B. Huang, K. Zhang, Y. Lin, B. Scholkopf, and C. Glymour, "Generalized score functions for causal discovery," *Knowledge Discovery and Data Mining*, vol. 2018, pp. 1551–1560, 2018.

[19] P. Nandy, A. Hauser, and M. H. Maathuis, "High-dimensional consistency in score-based and hybrid structure learning," *Annals of Statistic*, pp. 3151–3183, 2018.

[20] D. Koller and N. Friedman, *Probabilistic graph models: principles and techniques*. MIT press, 2009.

[21] K. Liu, Y. Li, X. Hu, M. Lucu, and W. D. Widanage, "Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3767–3777, 2020.

[22] M. Girolami and S. Rogers, "Variational bayesian multinomial probit regression with gaussian process priors," *Neural Computation*, vol. 18, no. 8, pp. 1790–1817, 2006.

[23] N. Cristianini, J. Shawetaylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," *Neural Information Processing Systems*, pp. 367–373, 2001.

[24] T. Handhayani and J. Cussens, "Kernel-based approach to handle mixed data for inferring causal graphs." *arXiv: Learning*, 2019.

[25] R. Tarjan, "Depth-first search and linear graph algorithms," *SIAM Journal on Computing*, pp. 146–160, 1972.

[26] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.

[27] J. Peters and P. Buhlmann, "Structural intervention distance (sid) for evaluating causal graphs," *arXiv: Machine Learning*, 2013.

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

[29] J. M. Mooij, J. Peters, D. Janzing, and et.al., "Distinguishing cause from effect using observational data: Methods and benchmarks," *Journal of Machine Learning Research*, vol. 17, no. 32, pp. 1–102, 2016.

[30] D. Margaritis, "Distribution-free learning of bayesian network structure in continuous domains," *American Association for Artificial Intelligence*, pp. 825–830, 2005.

[31] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.