# A Hybrid Causal Structure Learning Algorithm for Mixed-type Data

**Yan Li***, **Rui Xia***, **Chunchen Liu**, **Liang Sun**
DAMO Academy, Alibaba Group, Hangzhou, China
{yl.yy6993, guangyao.xr, chencang.lcc, liang.sun}@alibaba-inc.com

## Abstract

Inferring the causal structure of a set of random variables is a crucial problem in many disciplines of science. Over the past two decades, various approaches have been proposed for causal discovery from observational data. However, most of the existing methods are designed for either purely discrete or continuous data, which limit their practical usage. In this paper, we target the problem of causal structure learning from observational mixed-type data. Although there are a few methods that are able to handle mixed-type data, they suffer from restrictions, such as linear assumption and poor scalability. To overcome these weaknesses, we formulate the causal mechanisms via mixed structure equation model and prove its identifiability under mild conditions. A novel locally consistent score, named CVMIC, is proposed for causal directed acyclic graph (DAG) structure learning. Moreover, we propose an efficient conditional independence test, named MRCIT, for mixed-type data, which is used in causal skeleton learning and final pruning to further improve the computational efficiency and precision of our model. Experimental results on both synthetic and real-world data demonstrate that our proposed hybrid model outperforms the other state-of-the-art methods. Our source code is available at https://github.com/DAMO-DI-ML/AAAI2022-HCM.

## 1 Introduction

Discovering the underlying causal relations among multiple variables is beneficial in many applications, such as gene regulatory network reconstruction [20], understanding climate changes [12], and quantum analysis [31]. Conduction of experiments for causal relation identification [14] is usually expensive, time consuming and even unethical. Therefore, discovering the causal relations from purely observational data, commonly known as causal discovery or causal structure learning [15], is desired. In the real-world scenario, mixed-type data comprising both categorical and continuous variables is commonly observed, e.g., demographic attributes such as gender and occupation are often mixed with professional details including shopping behaviors in e-commerce. However, mixed-type data receives less attention in the causal discovery literature where techniques involving purely discrete or continuous variables have been advanced. The goal of this paper is to conduct causal discovery from i.i.d. samples with mixed-type variables whose distribution is Markov w.r.t. an underlying causal directed acyclic graph (DAG).

Structural equation models (SEMs) [21] and causal Bayesian network are the two major categories of causal discovery methods. SEMs formulate causal relations through specifying the equations between effects and causes, where additive noise models (ANM) [23, 5] and post-nonlinear [34] causal models have been widely and successfully applied for continuous variables. Peters et al. [22] propose ANM for processing discrete variables. Methods for causal Bayesian networks fall into two main categories: constraint-based and score-based ones. Constraint-based methods conduct conditional independence tests (CITs) to assess presence of edges. Log-likelihood ratio $G$-test and Pearson's $X^2$ are the typical CITs used on finite discrete variables. Pearson's correlation coefficient, kernel-based [35, 29] CITs and mutual information-based CITs [25] are the CITs commonly used for continuous variables. Score-based methods evaluate the quality of candidate causal structure using some properly defined score functions. Bayesian Information Criterion (BIC) [26] and Bayesian Dirichlet equivalent uniform (BDeu) [6] are the most commonly used score functions for continuous and discrete data, respectively. Recently in [36], Zheng et al. propose a continuous optimization framework for DAG structure learning and several extensions [37, 33] are built with the similar procedure. Unfortunately, all the aforementioned methods are designed for purely continuous or discrete data.

To handle mixed-type variables, traditional methods either discretize continuous variables [19, 11] or convert the conditional distributions of mixed-type variables into the same type [24]. More recently, the latent-LiNGAM proposed by Yamayoshi [32] introduces link functions to generate discrete variables. Copula PC [10] algorithm relaxes the rank-based measures of correlation in the continuous space to handle discrete variables. Causal MGM [27] uses likelihood ratio test to conduct CITs, where linear regression and Multinomial logistic regression are performed for continuous and categorical variables, respectively. The conditional Gaussian (CG) score [1] and degenerate Gaussian (DG) score [2] are proposed to adapt multivariate Gaussian model on mixed-type variables. The above-mentioned methods all rely on linear assumption of the causal mechanism that makes them restrictive. In [1], Andrews et al. propose the Mixed Variable

[*] These two authors contributed equally to this work.

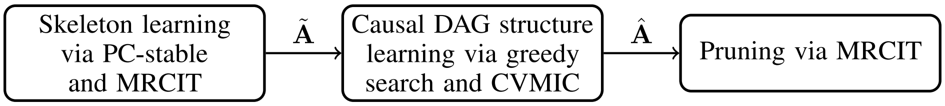
Skeleton learning via PC-stable and MRCIT
$\tilde{\mathbf{A}}$
Causal DAG structure learning via greedy search and CVMIC
$\hat{\mathbf{A}}$
Pruning via MRCIT

Figure 1: The framework of the proposed hybrid method. $\tilde{\mathbf{A}}$ is the output of the skeleton learning stage, i.e., a symmetric matrix to represent the adjacency matrix of an undirected graph. $\hat{\mathbf{A}}$ is the output of the second stage, i.e., adjacency matrix of a DAG.

Polynomial (MVP) score, which uses polynomial functions to model nonlinear causal relationships. The generalized score (GS) for causal discovery [17] can manipulate mixed-type and even multi-dimensional data while uncovering the nonlinear relations between variables. However, the MVP model has a strong assumption on the functional forms of causal mechanisms and the GS does not scale well as it requires regressions in the Reproducing Kernel Hilbert Space (RKHS).

To overcome the aforementioned limitations, we formulate the nonlinear causal mechanism for mixed-type data via a mixed-SEM, where causal mechanisms generating discrete variables are modeled as classifications and those generating continuous variables are formulated as nonlinear additive noise models (NAMs) [16]. We propose a generalized score function, **C**ross-**V**alidation based **M**ixed **I**nformation **C**riterion (CVMIC), to evaluate the quality of a candidate mixed-SEM. To efficiently search for the optimal mixed-SEM from the combinatorial DAG space, we develop a three-phase hybrid inference algorithm as illustrated in Figure 1: 1) A skeleton (i.e. edges without their orientations) is firstly learnt to reduce the search space, and we combine PC-stable algorithm [9] with our newly proposed CIT method that handles mixed-type data (named as **M**ixed-type **R**andomized CIT, MRCIT) to learn the skeleton structure. 2) Constrained by the skeleton structure, a greedy search procedure is then applied for DAG structure learning that starts with an empty DAG and greedily adds the edge corresponding to the *largest gain* in CVMIC score. 3) MRCIT is used again to prune the learned causal structure with a relatively larger conditional set aiming at reducing false positives. The proposed algorithm is named as "HCM" abbreviated for "**H**ybrid **C**ausal discovery on **M**ixed-type data".

The main contributions of our work are:

- We propose a mixed-SEM to formulate the nonlinear causal mechanism on mixed-type data, and prove its identifiability in the bivariate case.
- We propose an efficient conditional independence test for mixed-type data, named MRCIT.
- We propose a locally consistent information criterion, i.e., CVMIC, for causal DAG structure learning.

The rest of this paper is organized as follows: In section 2, we introduce the mixed-SEM and provide formal proof of its identifiability. The hybrid learning framework along with the details of the designed MRCIT and CVMIC are discussed in Section 3. In section 4, we elaborate experimental analysis on both synthetic and real-world data sets. And section 5 concludes our work.

## 2 Mixed Structural Equation Model

In this paper, a lowercase letter (e.g., $x$) denotes a specific value of a corresponding random variable (e.g., $X$), bold lowercase letters denote vectors (e.g., $\boldsymbol{x}$ ), and calligraphic uppercase letters signify sets (e.g., $\mathcal{X}$). $\mathbf{X}$, $\boldsymbol{x}_i$, $\boldsymbol{x}_{*,j}$, and $x_{i,j}$ represent the observations of all instances, observation of the $i$-th instance, $j$-th feature of all observed instances, and the $j$-th feature of the $i$-th observed instance, respectively.

Given $n$ random variables, i.e., $\mathcal{X} = \{X_1, X_2, \cdots, X_n\}$, the causal relations among them can be organized as a DAG $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where each $X_i$ can be either categorical ($X_i \in \{1, \cdots, c_i\}$) or continuous ($X_i \in \mathbb{R}$). $\mathcal{V} = \{1, 2, \cdots, n\}$ is a set of $n$ vertices/nodes, and each node is corresponding to a variable in $\mathcal{X}$. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of directed edges, and $j \to i$ is an edge representing that variable $X_j$ is a direct cause of variable $X_i$. $\mathcal{P}^{\mathcal{G}}_{(i)}$ denotes the index set of parent variables for $X_i$ given the causal graph $\mathcal{G}$, and note that $\mathcal{P}^{\mathcal{G}}_{(i)}$ can be $\emptyset$ for some $X_i$. A general SEM is defined as:

$$\begin{gathered}\{X_j = f_j(X_{\mathcal{P}^{\mathcal{G}}_{(j)}}, E_j) | j = 1, \cdots, n\},\ E_j \perp\!\!\!\perp X_{\mathcal{P}^{\mathcal{G}}_{(j)}} \\ \text{and } E_1, \cdots, E_n \text{ are mutually independent}, \qquad (1)\end{gathered}$$

where each variable is generated as a function of its directed causes and some noises, with causes and noises being independent. In this paper, we formulate a mixed-SEM to encode causal mechanism of mixed-type data. When $X_j$ is continuous, the nonlinear additive noise model (NAM) [16] $X_j = f_j(X_{\mathcal{P}^{\mathcal{G}}_{(j)}}) + E_j$ is used to formulate its generation. Therefore, its conditional distribution w.r.t. $f_j$ is:

$$\begin{aligned}\Pr(x_{i,j}|X_{\mathcal{P}^{\mathcal{G}}_{(j)}} = \boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}}, f_j) &= \Pr\Big(E_j = x_{i,j} - f_j(\boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}}) | X_{\mathcal{P}^{\mathcal{G}}_{(j)}}\Big) \\ &= \Pr\Big(E_j = x_{i,j} - f_j(\boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}})\Big), \quad (2)\end{aligned}$$

where the distributions of the noises $\{E_j | j = 1, \cdots, n\}$ are not presumed so that the proposed model has high generalization ability. When $X_j$ is discrete, it is generated via:

$$X_j = \mathop{\arg\max}_{k \in \{1, \cdots, c_j\}} f_{j,k}\left(X_{\mathcal{P}^{\mathcal{G}}_{(j)}}\right) + E_{j,k}, \qquad (3)$$

where $f_{j,k}(\cdot)$ is a numerical function of the $k$-th category of $X_j$, and $X_j$ has $c_j$ unique values in total. Its conditional distribution is realized via softmax function when noise term $E_{j,k}$ follows Gumbel distribution:

$$\Pr(X_j = l | X_{\mathcal{P}^{\mathcal{G}}_{(j)}} = \boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}}, f_j) = \frac{\exp\Big(f_{j,l}(\boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}})\Big)}{\sum_{c=1}^{c_j} \exp\Big(f_{j,c}(\boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}})\Big)}. \quad (4)$$

Given the causal graph $\mathcal{G}$, under the causal Markov condition, the joint distribution $\Pr(\mathcal{X} = \boldsymbol{x}_i)$ can be decomposed as the product of conditional distributions,

$$\Pr(\mathcal{X} = \boldsymbol{x}_i) = \prod_{j=1}^{n} \Pr(X_j = x_{i,j} | X_{\mathcal{P}^{\mathcal{G}}_{(j)}} = \boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}}), \quad (5)$$

Note that, $f_j(\cdot)$ in Eq.(2) and $f_{j,k}(\cdot)$ in Eq.(3) are nonlinear functions and the function classes of them are not specified. In our implementation, the LightGBM [18] is employed to model the causal mechanisms to achieve high generalisability. In addition, in LightGBM multiple regularization (e.g., L1 and L2 norms) and constraint (e.g., depth of tree) are introduced in the process of tree construction and hence Eq.(4) does not suffer from non-identifiable w.r.t. its parameters as the vanilla multi-logit model.

## 2.1 Identifiability of the Mixed-SEM

Identifiability is an essential requirement of causal models, and the formal definition of identifiability is given as follows:

**Definition 1** (Identifiability)**.** *Given data distribution $\Pr(\mathcal{X})$ generated by a SEM with graph $\mathcal{G}$, in particular, $\Pr(\mathcal{X})$ is Markov w.r.t. $\mathcal{G}$. We call $\mathcal{G}$* ***identifiable*** *from $\Pr(\mathcal{X})$ if $\Pr(\mathcal{X})$ cannot be generated by a SEM with a different graph $\mathcal{G}' \neq \mathcal{G}$.*

Based on [23, Remark 30] we conclude that the identifiability of a SEM in the bivariate case can be straightforward generalized under mild assumptions to the one in the multivariate case. Thus, here we focus on studying the identifiability of the above mixed-SEM in the bivariate case. In general, there are three scenarios in a bivariate mixed-SEM: 1) both variables are continuous, 2) both variables are categorical, 3) one categorical variable and one continuous variable. *In the first case, our model reduces to the standard NAM and its bivariate identifiability has been well studied in an existing work [16, Theorem 1], which indicates the bivariate NAM is not identifiable only if the differential equation for $\log p_x$ has a 3-dimensional space of solutions*. Next, we analyze the identifiability in the remaining two cases. The identifiability of SEM in bivariate case is commonly presented by showing the condition of non-identifiablity of $X \to Y$ and $Y \to X$ is restrictive and hard to satisfy in general [23], and we follows this procedure to present our results in the following.

**Lemma 1.** *Assuming variables $X$ and $Y$ are categorical and have finite support $X \in \{k_1, \dots, k_x\}$ and $Y \in \{t_1, \dots, t_y\}$. Let $\Pr(X, Y)$ admits a mixed-SEM from $Y$ to $X$ ($Y \to X$):*

$$X = \underset{k \in \{k_1, \dots, k_x\}}{\arg\max} \; f_k(Y) + E_k, \; E_k \perp\!\!\!\perp Y, \quad \text{and } E_k \sim \text{Gumbel distribution}.$$

*If $\Pr(X,Y)$ also allows a mixed-SEM from $X$ to $Y$ ($X \to Y$):*

$$Y = \underset{t \in \{t_1, \dots, t_y\}}{\arg\max} \; g_t(X) + E_t, \; E_t \perp\!\!\!\perp X, \quad \text{and } E_t \sim \text{Gumbel distribution},$$

*then for any quadruple $(k_i, k_j, t_a, t_b)$, where $(k_i, k_j) \in$ Supp$(X)$ and $(t_a, t_b) \in$ Supp$(Y)$, functions* **f** *and* **g** *must satisfy*

$$\begin{aligned} &g_{t_a}(k_i) + g_{t_b}(k_j) + f_{k_i}(t_b) + f_{k_j}(t_a) \\ =&g_{t_a}(k_j) + g_{t_b}(k_i) + f_{k_i}(t_a) + f_{k_j}(t_b). \end{aligned} \tag{6}$$

The formal proof of Lemma 1 can be found in Appendix B1.

Let us define a function $T(X, Y) = g_Y(X) - f_X(Y)$ over $X$ and $Y$, Lemma 1 indicates that, for any $(k_i, k_j) \in$ *Supp*$(X)$ and $(t_a, t_b) \in$ *Supp*$(Y)$, the four points, i.e.,

$$\{(k_i, t_a, T(k_i, t_a)), (k_j, t_b, T(k_j, t_b)), \\ (k_i, t_b, T(k_i, t_b)), (k_j, t_a, T(k_j, t_a))\},$$

are coplanar in a three-dimentional space, which is hard to achieve in general.

**Corollary 1 (Identifiability of Two Categorical Variables).** *Lemma 1 shows that $\Pr(X, Y)$ for two categorical variables admitting mixed-SEMs in both directions falls into a* restrictive space*, which is hard to be satisfied in general. Therefore,* ***in general, bivariate mixed-SEM is identifiable when both variables are categorical.***

**Lemma 2.** *Assuming variable $X$ is categorical that has finite support (i.e., $X \in \{k_1, \dots, k_x\}$), and $Y$ is a continuous variable. Let $\Pr(X, Y)$ admit a mixed-SEM from $Y$ to $X$ ($Y \to X$):*

$$X = \underset{k \in \{k_1, \dots, k_x\}}{\arg\max} \; f_k(Y) + E_k.$$

*If $\Pr(X, Y)$ also allows a mixed-SEM from $X$ to $Y$ ($X \to Y$), with the additiveness of noise relaxed as $Y = g(X, E_Y)$, where $g$ is invertible and $E_Y \perp\!\!\!\perp X$. Then the following condition must be satisfied*

$$\sum_x P(X = x) \left[ \frac{p_{E_y}(g_x^{-1}(y)) p'_y(y)}{p_y(y)} - \frac{p'_{E_y}(g_x^{-1}(y))}{g'_x(g_x^{-1}(y))} \right] = 0,$$

*where $p_y(y)$, $p'_y(y)$, $p_{E_y}(\cdot)$ and $p'_{E_y}(\cdot)$ denote the probability density functions and their gradient w.r.t. $y$ and $E_y$, respectively. Note that we have rewritten the mixed-SEM from $X$ to $Y$ as $Y = g_X(E_Y)$, since $X$ is categorical and we have $|X|$ functions, i.e. $\{g_{k_1}(E_Y), \dots, g_{k_x}(E_Y)\}$.*

The formal proof of Lemma 2 can be found in Appendix B2.

**Corollary 2 (Identifiability of One Categorical and One Continuous Variable).** *Lemma 2 shows that $\Pr(X, Y)$ for one categorical and one continuous variables admitting mixed-SEMs in both directions is hard to be satisfied in general. Therefore,* ***in general, bivariate mixed-SEM is identifiable when one variable is categorical and the other one is continous.***

It is worth mentioning that [23, Definition 27] is necessary for adapting the identifiability from the bivariate case to the multivariate case in our model. Let us rewrite [23, Definition 27] for mixed-type data. In multivariate case, our proposed mixed-SEM is identifiable under the following condition:

For all $j \in \mathcal{V}$, $i \in \mathcal{P}^{\mathcal{G}}_{(j)}$ and all sets $\mathcal{S} \subseteq \mathcal{V}$ with $\mathcal{P}^{\mathcal{G}}_{(j)} \backslash i \subseteq \mathcal{S} \subseteq ND_j \backslash \{i, j\}$ (i.e., $ND_j$ represents the non-descendents of $j$), there is $X_{\mathcal{S}} = x_{\mathcal{S}}$, with $p_{\mathcal{S}}(x_{\mathcal{S}}) > 0$, s.t., 1) When both $X_i$ and $X_j$ are categorical, $f_{X_j}(X_i|_{X_\mathbf{S} = x_\mathbf{S}}, X_\mathbf{S} = x_\mathbf{S})$ and $g_{X_i}(X_j|_{X_\mathbf{S} = x_\mathbf{S}}, X_\mathbf{S} = x_\mathbf{S})$ do not satisfy Eq.(6) for any quadraple; 2) When one of them (e.g., $Y_i$) is continuous, then $p_{Y_i}(Y_j|_{Y_\mathbf{S} = y_\mathbf{S}}, Y_\mathbf{S} = y_\mathbf{S})$, $p_{E_{Y_i}}$, and their gradients do not satisfy the nonidentifiable condition in Lemma 2; 3) When both $X_i$ and $X_j$ are continuous, then $(g_{X_i}(X_j|_{X_\mathbf{S} = x_\mathbf{S}}, X_\mathbf{S} = x_\mathbf{S}), E_{X_i})$ do not satisfy the nonidentifiable condition [23, Condition 19].

The above statements can be proved directly based on the proof of [23, Theorem 28], since it does not require model assumptions in the proof.

# 3 Hybrid Structure Learning Algorithm

To efficiently infer the mixed-SEM from data, we propose a three-phase hybrid learning framework as shown in Figure 1. We first introduce the proposed CVMIC and MRCIT, and then provide implementation details.

## 3.1 CVMIC: Cross-Validation based Mixed Information Criterion

Score-based causal inference algorithms conduct causal structure learning by optimizing a properly defined score function. Since the maximized likelihood may lead to overfitting in structure learning, we define a CVMIC score to measure the fitness of a candidate mixed-SEM on mixed-type data. Cross-validation (CV) is a widespread strategy for model selection in machine learning and statistics [3], which evaluates the model's ability to handle new data. We utilize CV to identify causal structures since causal relationship is considered more stable than correlation and hence more robust to unobserved instances [4]. Without loss of generality, we use k-fold CV. In the $q$-th round, the log-likelihood of random variable $X_j$ evaluated on the $q$-th testing set ($\mathbf{X}_{\mathcal{D}_q}$) is formulated as:

$$\begin{aligned}&L(X_j = \boldsymbol{x}_{\mathcal{D}_q,j}|\mathcal{P}^{\mathcal{G}}_{(j)}; \hat{f}^{(q)}_j)\\ =&\sum_{i\in\mathcal{D}_q} \log\left(\Pr(X_j = x_{i,j}|X_{\mathcal{P}^{\mathcal{G}}_{(j)}} = \boldsymbol{x}_{i,\mathcal{P}^{\mathcal{G}}_{(j)}}; \hat{f}^{(q)}_j)\right),\end{aligned}$$

where $\mathcal{D}_q$ is the index set of the $q$-th testing set, and $\hat{f}^{(q)}_j$ is the maximum likelihood estimation (MLE) of the causal mechanism learned from the $q$-th training set. The cross-validated log-likelihood of $X_j$ can be calculated as $\sum_{q=1}^{k} L(X_j = \boldsymbol{x}_{\mathcal{D}_q,j}|\mathcal{P}^{\mathcal{G}}_{(j)}; \hat{f}^{(q)}_j)$, and inspired by BIC we propose the final score function for the $j$-th random variable:

$$S_j(\mathcal{P}^{\mathcal{G}}_{(j)};X_j)=\sum_{q=1}^{k} L(X_j=\boldsymbol{x}_{\mathcal{D}_q,j}|\mathcal{P}^{\mathcal{G}}_{(j)}; \hat{f}^{(q)}_j)-\frac{\log(m)}{2}|\mathcal{P}^{\mathcal{G}}_{(j)}|,$$

where $m$ is the sample size. For $n$ random mixed-type variables, our proposed score function CVMIC is defined as:

$$S(\mathcal{G};\mathbf{X}) = \sum_{j=1}^{n} S_j(\mathcal{P}^{\mathcal{G}}_{(j)}; X_j) \tag{7}$$

To theoretically evaluate the soundness of our proposed CVMIC score, we give the definition of local consistency and prove that CVMIC score is locally consistent. Generally speaking, local consistency means that optimizing the model selection criterion leads to selecting an edge does not conflict with any independence constraint [8].

**Definition 2 (Local Consistency).** *Let $\mathbf{X}$ be $m$ i.i.d. samples drawn form a distribution $P(\cdot)$, $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG resulted from adding the edge $i \to j$ in $\mathcal{G}$. A scoring criterion $S(\mathcal{G};\mathbf{X})$ is locally consistent if the following two properties hold as the sample size $m \to \infty$:*

1. *If $X_i \not\perp\!\!\!\perp X_j|X_{\mathcal{P}^{\mathcal{G}}_{(j)}}$, then $S(\mathcal{G}';X) > S(\mathcal{G};\mathbf{X})$.*
2. *If $X_i \perp\!\!\!\perp X_j|X_{\mathcal{P}^{\mathcal{G}}_{(j)}}$, then $S(\mathcal{G}';X) < S(\mathcal{G};\mathbf{X})$.*

**Theorem 1.** *CVMIC score is locally consistent.*

The proof of Theorem 1 can be found in Appendix B3.

When the number of samples goes to infinity, the log-likelihood is globally consistent, i.e., it reaches the minimum when the true graph structure is estimated. In CVMIC, a penalty term is added to cross-validated log-likelihood to encourage each node to have less parents and hence introduce bias and fail to reach minimum when the true graph is estimated. Since the sample size is limited in practice, the CVMIC works better than the vanilla log-likelihood even it is not globally consistent. Moreover, the structure learning is a NP-hard problem which is not practical to conduct exhaustive search. Therefore, in structure learning local consistency has more practical value than global consistency. In Table 1 within Section 4.2 we have included the method "HCM-ll", where the log-likelihood is used as score function, and the results demonstrate the advantage of using CVMIC.

## 3.2 MRCIT: A Randomized Conditional Independence Test for Mixed-type Data

As mentioned in Introduction section and Figure 1, in our hybrid algorithm, CITs are used in both causal skeleton learning and final pruning to reduce the search space of causal DAG learning and reduce false positive, respectively. Therefore, we propose a randomized CIT for Mixed-type Data. To test whether $X_1 \perp\!\!\!\perp X_2|\mathbf{X}_3$, where $\mathbf{X}_3$ can be multivariate, a nonparametric methods called KCIT is introduced by Zhang [35]. Strobl et al. [29] further propose RCIT utilizing random Fourier features (RFFs) to get a faster CIT without sacrificing accuracy. However, cases when variables $\{X_1, X_2, \mathbf{X}_3\}$ involving both continuous and categorical variables are not considered. Here we extend RCIT to handle mixed-type data.

Let $\mathcal{H}_{\mathcal{X}}$ be a Hilbert space of function mapping from $\mathcal{X}$ to $\mathbb{R}$, the partial cross-covariance operator is related with CITs as $\|\Sigma_{\ddot{X}X_2\cdot\mathbf{X}_3}\|^2_{HS} = 0 \iff \Sigma_{\ddot{X}X_2\cdot\mathbf{X}_3} = 0 \iff X_1 \perp\!\!\!\perp X_2|\mathbf{X}_3$, where $\ddot{X} = (X_1, \mathbf{X}_3)$. CIT is then relaxed to test for uncorrelatedness between functions in Hilbert spaces, i.e., any residual function of $(X_1, \mathbf{X}_3)$ given $\mathbf{X}_3$ is uncorrelated with that of $(X_2, \mathbf{X}_3)$ given $\mathbf{X}_3$ if and only if $X_1 \perp\!\!\!\perp X_2|\mathbf{X}_3$. KCITs choose to directly regress out the effect of the conditional set $\mathbf{X}_3$ through kernel ridge regression (KRR), where inverse of kernel $K_{\mathbf{X}_3}$ is required that scales cubically with sample size. To avoid such inversion, Frobenius norm corresponding to Hilbert-Schmidt norm in Euclidean space of $\mathcal{C}_{\ddot{A}B\cdot\mathbf{X}_3} = \mathbb{E}[(\ddot{A}_i - \mathbb{E}(\ddot{A}|\mathbf{X}_3))(B_i - \mathbb{E}(B|\mathbf{X}_3))]$ is used. $\ddot{A}$ and $B$ are RFFs of $\ddot{X}$ and $X_2$ respectively, which help to approximate continuous shift-invariant kernels using the following result:

$$k(x, y) = \int e^{iw^T(x-y)}dF_w = \mathbb{E}[\varphi(x)\varphi(y)],$$

where $\varphi(x) = \sqrt{2}\cos(W^T x + B), W \sim \mathbb{P}_W, B \sim \text{Uniform}([0, 2\pi])$. When estimating non-linear functions of $\mathbf{X}_3$ for $\mathbb{E}(\ddot{A}|\mathbf{X}_3)$, the original KRR problem is converted to $\hat{f}(\mathbf{X}_3) = \mathbf{K}(\mathbf{X}_3, \mathbf{X}_3)\boldsymbol{\alpha} = \boldsymbol{\varphi}(\mathbf{X}_3)^*\mathbf{w}$, where solving $\mathbf{w}$ requires inverting a much smaller matrix than that for solving $\boldsymbol{\alpha}$. Considering $\mathbf{X}_3$ might contain both categorical and continuous variables in our case, it is improper to apply Gaussian

Kernel for all variables since it might lead to loss of information. We use the following formulation to separate kernels for continuous and categorical variables, and also derive the corresponding RFFs when estimating the separated kernels. Denote $\mathbf{x} = [\mathbf{x}_{ct}, \mathbf{x}_{ca}]$, with $\mathbf{x}_{ct}$ for continuous and $\mathbf{x}_{ca}$ for categorical variables, respectively.

$$
\begin{aligned}
&k(\mathbf{x}, \mathbf{y}) = \int e^{i\omega_{ct}^T \Delta_{ct}} dF_{\omega_{ct}} + \int e^{i\omega_{ca}^T \Delta_{ca}} dF_{\omega_{ca}} \approx \frac{1}{S} \mathrm{Tr}(\varphi(\mathbf{x})\varphi(\mathbf{y})^*) \\
&\sum_{j=1}^{n} k(\mathbf{x}_j, \mathbf{x})\alpha_j \\
&= \sum_{j=1}^{n} \frac{1}{S} \sum_{s} \Big[ \varphi_{ct}(\mathbf{x}_{jct})^* \varphi_{ct}(\mathbf{x}_{ct}) + \varphi_{ca}(\mathbf{x}_{jca})^* \varphi_{ca}(\mathbf{x}_{ca}) \Big] \alpha_j \\
&= \varphi_{ct}(\mathbf{x}_{ct})^* \mathbf{w}_{ct} + \varphi_{ca}(\mathbf{x}_{ca})^* \mathbf{w}_{ca}, \qquad (8)
\end{aligned}
$$

where $S$ is the number of RFFs and $\varphi(\mathbf{x})$ is the concatenation of continuous and categorical RFFs. During the experiment, twice of median distance between points in the input space is set as the kernel width for continuous variables, while for categorical variable, tiny kernel width is chosen to approximate the Dirac function.

### 3.3 Implementation

**Skeleton Learning** Under the causal Markov assumption and causal faithfulness assumption [28], we employ PC-stable [9] that iteratively enlarges the size of the conditional set and discard edges between nodes if their corresponding variables are (conditional) independent via MRCIT. To accelerate our algorithm, MRCITs are parallelly tested within each iteration and the maximum size of the conditional set is set to 1 in our implementation. Larger conditional sets are not advised as we wish to identify cause and effect before discarding potential edges that would lead to false negatives if co-children are also considered into the conditional sets.

**Causal DAG Structure Learning** Since the value space of a discrete variable is limited, the estimated log-likelihood of a classification model is usually larger than that of a regression model. As a result, for standard search algorithms, such as hill climbing [13], adding an edge with the largest score may lead to systematic selection bias, i.e., preference of selecting an edge directed to a discrete variable. To alleviate this, we introduce a greedy procedure (summarized as Algorithm 1 in Appendix A) that starts with an empty DAG and adds the edge corresponding to the ***largest gain*** in score during each iteration. The initial score of the $j$-th variable is defined as: $S(\emptyset, X_j) = \frac{1}{n}\sum_{i=1}^{n} \log(\Pr(X_j = x_{i,j}))$, where $\Pr(X_j = x_{i,j})$ is the empirical or observational probability of $X_j = x_{i,j}$. The empirical probability for a discrete variable is calculated via frequency, while for a continuous variable, it is estimated via kernel density estimation (KDE). In each iteration, the gain of adding a potential edge $l \to j$ is computed as $S_j(\mathcal{P}_{(j)}^{\mathcal{G}} \cup \{l\}; X_j) - S_j(\mathcal{P}_{(j)}^{\mathcal{G}}; X_j)$, and the edge with largest gain will be added to the DAG. To assure acyclic, a directed-path-matrix ($\mathbf{O}$) will be updated accordingly, and an edge $l \to j$ will not be considered as a valid candidate if $o_{j,l} = 1$, i.e, there is a path from node $j$ to node $l$. Note that, the score of the $j$-th node and the gain of adding a potential edge is updated only if the parents set of the $j$-th node is updated and hence being efficient. For a continuous variable, the likelihood in CVMIC score is also estimated via KDE to avoid assuming the noise distribution. The soundness of using KDE in likelihood computation is proved in [7]. We employ LightGBM [18] in this work for non-parametric estimation of the causal mechanism. Moreover, the search space has been reduced after skeleton learning and the algorithm is feasible for up to hundreds of nodes.

**Pruning** In the previous stage, an edge will be greedily added as long as it does not break the DAG constraint. This setting helps us to avoid choosing an intractable threshold for determining the minimum score gain of adding an edge. However, this setting tends to add more "superfluous" edges and pruning is required to improve precision of the inference. Pruning is conducted via MRCITs to test independence of each parent-child pair conditioning on all the other direct causes of the child, and remove "superfluous" edges if the corresponding parent and child are conditional independent.

## 4 Experimental Result

### 4.1 Comparison methods and experiment setup

We compare our proposed HCM algorithm with six related state-of-the-art methods, which are capable of conducting causal discovery on mixed-type data. These methods are: Copula PC (CPC) [10] [1], Causal MGM (MGM) [27] [2], Conditional Gaussian (CG) score [1] [3], Degenerate Gaussian (DG) score [2] [3], Mixed Variable Polynomial (MVP) score [1] [3], and the generalized score (GS) for causal discovery [17] [4]. The last two competed models are able to handle nonlinear causal mechanisms. We also compare our model with a hybrid method, i.e., Max-Min Hill-Climbing (MMHC) algorithm [30] [5], by discretizing the continuous variable. In all experiments, thresholds of the $p$-value in CITs are set to $0.05$. We use the default kernel functions in GS. Regularization parameters in MGM are tuned and the best result in each synthetic data is reported. The remaining comparison methods, i.e, DG, CG, MVP, do not have complex hyper-parameters to tune and hence all corresponding experiments are conducted with default settings. All of our experiments are conducted on a machine with 2.6 GHz 6-Core Intel i7 CPU and 16GB DDR4 2667 MHz RAM.

### 4.2 Synthetic data generated from benchmark simulator

Real data sets with mixed-type variables and ground truth of causal DAG structures are very hard to find. Therefore, we use a benchmark simulator in *Tetrad* [3] to generate six data sets based on DAGs with different numbers of nodes

[1] https://cran.r-project.org/web/packages/copula/copula.pdf

[2] http://causalmgm.org

[3] https://github.com/cmu-phil/tetrad

[4] https://github.com/Biwei-Huang/Generalized-Score-Functions-for-Causal-Discovery

[5] https://www.bnlearn.com/documentation/man/structure.learning.html

($n = 50, 100$) and different average node degrees (3, 10, 20). Note that the graph is more dense when the node degree (the number of edges connected to the node) is larger. The number of instances in these data sets are set to be $100 \times n$, and the percentage of discrete variables are set to be $50\%$. In this simulator, discrete variables are generated via randomly parameterized Multinomial distribution.

Table 1: Performance comparison of HCM against the state-of-the-art methods w.r.t. Precision (Prec.), Recall (Rec.), F1-score (F1), and Structural hamming distance (SHD) on the synthetic data sets generated from simulator in Tetrad.

| Avg Deg. | Method | # of nodes = 50 | | | | # of nodes = 100 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Prec. | Rec. | F1 | SHD | Prec. | Rec. | F1 | SHD |
| 3 | HCM | 0.600 | **0.640** | **0.619** | **35** | 0.610 | **0.700** | **0.652** | **77** |
| | HCM-RCIT | 0.597 | 0.533 | 0.563 | 39 | 0.454 | 0.393 | 0.421 | 107 |
| | HCM-ll | 0.436 | 0.453 | 0.444 | 49 | 0.517 | 0.593 | 0.553 | 94 |
| | HCM-nocv | 0.422 | 0.467 | 0.443 | 51 | 0.506 | 0.567 | 0.535 | 94 |
| | MVP | 0.800 | 0.267 | 0.400 | 56 | 0.793 | 0.433 | 0.560 | 92 |
| | CG | 0.536 | 0.400 | 0.458 | 60 | **0.795** | 0.440 | 0.567 | 93 |
| | DG | 0.548 | 0.453 | 0.496 | 51 | 0.687 | 0.527 | 0.596 | 83 |
| | MGM | 0.500 | 0.093 | 0.157 | 64 | 0.492 | 0.200 | 0.284 | 102 |
| | CPC | 0.451 | 0.427 | 0.438 | 85 | 0.449 | 0.233 | 0.307 | 275 |
| | MMHC | **0.846** | 0.440 | 0.579 | 44 | 0.733 | 0.440 | 0.550 | 91 |
| 10 | HCM | 0.566 | **0.296** | **0.388** | **156** | 0.615 | **0.385** | **0.474** | **289** |
| | HCM-RCIT | 0.474 | 0.133 | 0.208 | 181 | 0.593 | 0.234 | 0.336 | 324 |
| | HCM-ll | 0.543 | 0.281 | 0.370 | 162 | 0.442 | 0.278 | 0.341 | 331 |
| | HCM-nocv | 0.491 | 0.256 | 0.337 | 168 | 0.415 | 0.268 | 0.326 | 345 |
| | MVP | 0.800 | 0.099 | 0.175 | 184 | 0.756 | 0.166 | 0.272 | 349 |
| | CG | 0.889 | 0.158 | 0.268 | 173 | 0.814 | 0.234 | 0.364 | 324 |
| | DG | 0.750 | 0.148 | 0.247 | 178 | 0.674 | 0.217 | 0.328 | 331 |
| | MGM | 0.632 | 0.059 | 0.108 | 187 | 0.505 | 0.112 | 0.184 | 331 |
| | CPC | 0.535 | 0.227 | 0.318 | 195 | 0.509 | 0.137 | 0.215 | 447 |
| | MMHC | **0.907** | 0.192 | 0.317 | 166 | **0.851** | 0.237 | 0.370 | 317 |
| 20 | HCM | 0.486 | 0.176 | 0.258 | 347 | 0.676 | **0.335** | **0.448** | **309** |
| | HCM-RCIT | 0.451 | 0.119 | 0.188 | 353 | 0.514 | 0.174 | 0.260 | 374 |
| | HCM-ll | 0.308 | 0.127 | 0.179 | 380 | 0.438 | 0.223 | 0.296 | 364 |
| | HCM-nocv | 0.335 | 0.140 | 0.197 | 372 | 0.425 | 0.219 | 0.289 | 363 |
| | MVP | 0.593 | 0.090 | 0.157 | 357 | 0.755 | 0.093 | 0.166 | 390 |
| | CG | 0.464 | 0.116 | 0.186 | 357 | 0.667 | 0.130 | 0.218 | 384 |
| | DG | 0.438 | 0.109 | 0.174 | 362 | 0.596 | 0.123 | 0.204 | 380 |
| | MGM | 0.458 | 0.028 | 0.054 | 369 | 0.480 | 0.059 | 0.104 | 371 |
| | CPC | 0.510 | 0.083 | 0.142 | 347 | 0.473 | 0.081 | 0.139 | 516 |
| | MMHC | **0.762** | **0.207** | **0.325** | **314** | **0.890** | 0.170 | 0.285 | 357 |

Higher F1 score and lower SHD indicate better performance. From results in Table 1 we can observe that our proposed model outperforms the other methods in most data sets. Note that, the GS does not scale to these six data sets and hence the results are not included. In addition, when the graph is dense, the maximal number of parents of a node can be as large as 10, making the probability table of Multinomial distribution very large and hard to estimate for all methods. Comparing with the model performance of HCM and "HCM-nocv", we demonstrate the advantage of using cross-validation in score function. Comparing with HCM and "HCM-RCIT", where RCIT [29] is used for conditional independence test in step 1 and step 3, we demonstrate the advantage of our proposed MRCIT. The results in Table 1 also demonstrate the advantage of using CVMIC (score function in HCM) over log-likelihood (score function in "HCM-ll").

**Ablation study** Table 2 presents the results of ablation study of our HCM. "StablePC+MRCIT" takes MRCIT for CITs in PC-Stable algorithm [9] and employs the orientation rules in PC. Therefore, it is a constraint-based structure learning method based on our proposed MRCIT. In "StablePC+MRCIT", a large number (super exponential w.r.t. number of nodes) of CITs with large conditional sets are required and hence it can only handle relatively small/sparse graph. The running time also demonstrates that HCM scales better than "StablePC+MRCIT" in dense graph. "HCM_no_step3" represents the result after step 2 without pruning, and we can observe that step 3 in general can improve the result with relatively small time cost. The ablation study shows the strength of the hybrid structure in HCM.

### 4.3 Synthetic data generated from benchmark network structures

We also generate mixed-type synthetic data based on some benchmark network structures in causal discovery[6], which are summarized in Table 3.

Let us define the source variables are the ones without parents and the remaining ones are non-source. A continuous non-source variable $X_j$ is generated as $X_j = f_j(X^{\mathcal{G}}_{\mathcal{P}_{(j)}}) + E_j$, and a discrete non-source variable $X_l$ is generated as $X_l = \arg\max_{k \in \{1, \cdots, c_l\}} f_{l,k}(X^{\mathcal{G}}_{\mathcal{P}_{(l)}}) + E_{l,k}$. $E_j$ and $E_{l,k}$ are the noise terms randomly drawn from a set of noise distributions. $f_j$ and $f_{l,k}$ are the functional causal mechanisms. $c_l$ is the number of unique values in $X_l$, which is randomly chosen from 2 to 10. Continuous and discrete source variables are generated using the selected noise distribution and Multinomial distribution, respectively. For each network, we simulate the mixed-type data sets with:

- Various percentages of discrete variables: 0.2 and 0.5.
- Various noise distributions: Normal, Uniform, and Exponential.
- Various functional causal mechanisms: 1) “mixed-additive” function, i.e., a weighted summation of $x$, $x^2$, $\sin x$, $\sin(x^2)$, $\tan(x)$. 2) “modified-sigmoid” function, i.e., a weighted combination of $\frac{b \cdot (x+a)}{1+|b(x+a)|}$, where $a$ and $b$ are randomly chosen coefficients.

In summary, we generate 12 mixed-type data sets for each network, and the sample size of each data set is set to be $100 \times n$, where $n$ is the number of nodes.

In Figure 2, we present the averaged F1-score and normalized structural hamming distance (N-SHD) [30] to evaluate the performance of each causal discovery method. Higher F1 score and lower N-SHD indicate better performance. We can thus observe that our proposed model outperforms the other methods. Note that, as the generalized score method [17] does not scale very well, we are not able to get its result after 24 hours on larger graphs, i.e., HAILFINDER, HEPAR, and ANDES. Copula PC is also not scalable to ANDES under the same experimental setting.

[6] The typologies of networks is available in https://www.bnlearn.com/bnrepository/

Table 2: Ablation study of the proposed three-phase hybrid algorithm.

| # of Nodes | Avg Deg. | Method | Performance | | | | Execution time (s) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Pre. | Rec. | F1 | SHD | Total | Step 1 | Step 2 | Step 3 |
| 50 | 3 | HCM | 0.600 | 0.640 | 0.619 | 35 | 467.76 | 62.82 | 401.97 | 2.96 |
| | | HCM_no_step3 | 0.600 | 0.640 | 0.619 | 35 | 464.79 | 62.82 | 401.97 | NA |
| | | StablePC+MRCIT | 0.894 | 0.560 | 0.689 | 35 | 2577.37 | NA | NA | NA |
| | 10 | HCM | 0.566 | 0.296 | 0.388 | 156 | 1652.65 | 664.03 | 865.19 | 123.43 |
| | | HCM_no_step3 | 0.556 | 0.296 | 0.386 | 158 | 1529.22 | 664.03 | 865.19 | NA |
| | | StablePC+MRCIT | 0.667 | 0.099 | 0.172 | 186 | 3126.86 | NA | NA | NA |
| | 20 | HCM | 0.486 | 0.176 | 0.258 | 347 | 3314.51 | 1565.18 | 1353.93 | 395.41 |
| | | HCM_no_step3 | 0.467 | 0.181 | 0.261 | 353 | 2919.1 | 1565.18 | 1353.93 | NA |
| | | StablePC+MRCIT | 0.660 | 0.085 | 0.151 | 360 | 10416.45 | NA | NA | NA |
| 100 | 3 | HCM | 0.610 | 0.700 | 0.652 | 77 | 3196.59 | 1772.2 | 1381.3 | 43.09 |
| | | HCM_no_step3 | 0.610 | 0.700 | 0.652 | 77 | 3153.5 | 1772.2 | 1381.3 | NA |
| | | StablePC+MRCIT | 0.855 | 0.473 | 0.609 | 82 | 14938.66 | NA | NA | NA |
| | 10 | HCM | 0.615 | 0.385 | 0.474 | 289 | 4663.89 | 2635.24 | 1758.4 | 270.24 |
| | | HCM_no_step3 | 0.605 | 0.400 | 0.482 | 291 | 4393.64 | 2635.24 | 1758.4 | NA |
| | | StablePC+MRCIT | Fail to get result after 12 hours runing | | | | | | | |
| | 20 | HCM | 0.676 | 0.335 | 0.448 | 309 | 5212.29 | 2701.26 | 2111.96 | 399.07 |
| | | HCM_no_step3 | 0.676 | 0.340 | 0.340 | 308 | 4813.22 | 2701.26 | 2111.96 | NA |
| | | StablePC+MRCIT | Fail to get result after 12 hours runing | | | | | | | |

Table 3: Number of nodes and edges in each benchmark network structure.

| name | MILDEW | ALARM | HAILFINDER | HEPAR | ANDES |
|---|---|---|---|---|---|
| # of nodes | 35 | 37 | 56 | 70 | 223 |
| # of edges | 46 | 46 | 66 | 123 | 338 |

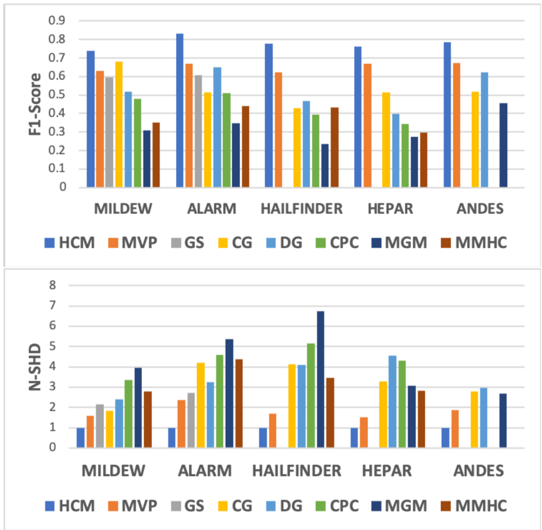

Figure 2: The model performance w.r.t. F1-score and N-SHD on synthetic data sets. The results are averaged over 12 synthetic data sets for each network.

## 4.4 Real-World Application

**Credit Data** We apply our method on the German credit data set[7], where the original goal is to classify people as good or bad credit risks based on the attributes of financial history and personal status. We wish to discover important attributes that directly cause the target variable, credit risk. The data set contains 21 variables, with 7 continuous and 14 categorical

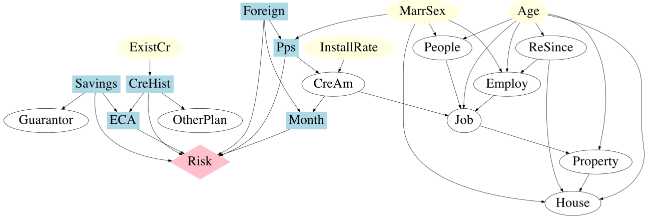

Figure 3: Learned causal graph from credit data set. The red diamond denotes the target, i.e., credit risks. Blue rectangles are direct causes of credit risks.

where continuous variables include credit amount(CreAm), installment rate in percentage of disposable income (InstallRate), and categorical variables include owned property, status of marriage and gender(MarrSex). Figure 3 shows the learned causal graph. The direct causes of credit risks are highlighted with blue rectangles, including savings bonds, credit history, status of existing checking account, duration in month, purpose and foreign, most of which are financial related and do not involve demographic information except being foreign. In addition, separation of financial and demographic attributes is clear in the whole graph, and the two parts communicate through credit purpose and duration length. The recovered causal relations are in accordance with our common understandings and domain knowledge.

## 5 Conclusion

In this paper, we propose a hybrid algorithm, named HCM, for causal structure learning on mixed-type data, i.e., data sets with both continuous and categorical variables. We propose a new score CVMIC for accurate causal DAG learning and a novel conditional independence test MRCIT on mixed-type data. We also theoretically analyze the identifiability and local consistency of our proposed model. For future work, we plan to further improve the computational efficiency of our approach and extend it to the cases where there are unobserved confounders in the underlying causal graph.

[7] Available at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

## References

[1] Andrews, B.; Ramsey, J.; and Cooper, G. F. 2018. Scoring Bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1): 3–18.

[2] Andrews, B.; Ramsey, J.; and Cooper, G. F. 2019. Learning high-dimensional directed acyclic graphs with mixed data-types. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, 4–21. PMLR.

[3] Arlot, S.; Celisse, A.; et al. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: 40–79.

[4] Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

[5] Bühlmann, P.; Peters, J.; Ernest, J.; et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of statistics*, 42(6): 2526–2556.

[6] Buntine, W. 1991. Theory refinement on Bayesian networks. In *Uncertainty Proceedings 1991*, 52–60. Elsevier.

[7] Cai, R.; Qiao, J.; Zhang, Z.; and Hao, Z. 2018. Self: structural equational likelihood framework for causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[8] Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554.

[9] Colombo, D.; and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1): 3741–3782.

[10] Cui, R.; Groot, P.; and Heskes, T. 2016. Copula PC algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 377–392. Springer.

[11] Dojer, N. 2016. Learning Bayesian networks from datasets joining continuous and discrete variables. *International Journal of Approximate Reasoning*, 78: 116–124.

[12] Ebert-Uphoff, I.; and Deng, Y. 2012. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17): 5648–5665.

[13] Gámez, J. A.; Mateo, J. L.; and Puerta, J. M. 2011. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22(1): 106–148.

[14] Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

[15] Heinze-Deml, C.; Maathuis, M. H.; and Meinshausen, N. 2018. Causal structure learning. *Annual Review of Statistics and Its Application*, 5: 371–391.

[16] Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21: 689–696.

[17] Huang, B.; Zhang, K.; Lin, Y.; Schölkopf, B.; and Glymour, C. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1551–1560.

[18] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30: 3146–3154.

[19] Monti, S.; and Cooper, G. 1998. A Multivariate Discretization Method for Learning Bayesian Networks from Mixed Data. 404–413.

[20] Neto, E. C.; Keller, M. P.; Attie, A. D.; and Yandell, B. S. 2010. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The annals of applied statistics*, 4(1): 320.

[21] Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge university press.

[22] Peters, J.; Janzing, D.; and Scholkopf, B. 2011. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450.

[23] Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal Discovery with Continuous Additive Noise Models. 15(1).

[24] Romero, V.; Rumí, R.; and Salmerón, A. 2006. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42(1-2): 54–68.

[25] Runge, J. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, 938–947. PMLR.

[26] Schwarz, G.; et al. 1978. Estimating the dimension of a model. *Annals of statistics*, 6(2): 461–464.

[27] Sedgewick, A. J.; Shi, I.; Donovan, R. M.; and Benos, P. V. 2016. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC bioinformatics*, 17(5): 307–318.

[28] Spirtes, P.; and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, 1–28. SpringerOpen.

[29] Strobl, E. V.; Zhang, K.; and Visweswaran, S. 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1).

[30] Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1): 31–78.

[31] Wood, C. J.; and Spekkens, R. W. 2015. The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning. *New Journal of Physics*, 17(3): 033002.

[32] Yamayoshi, M.; Tsuchida, J.; and Yadohisa, H. 2020. An estimation of causal structure based on Latent LiNGAM for mixed data. *Behaviormetrika*, 47(1): 105–121.

[33] Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, 7154–7163. PMLR.

[34] Zhang, K.; and Hyvarinen, A. 2012. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*.

[35] Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

[36] Zheng, X.; Aragam, B.; Ravikumar, P.; and Xing, E. P. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 9492–9503.

[37] Zheng, X.; Dan, C.; Aragam, B.; Ravikumar, P.; and Xing, E. 2020. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 3414–3425. PMLR.