Mixed Causal Structure Discovery with Application to Prescriptive Pricing

Wei Wenjuan[*], **Feng Lu**[*] and **Liu Chunchen**[†] [*]
NEC Labs China
wei_wenjuan@nec.cn, feng_lu@nec.cn, liu_chunchen@nec.cn

Abstract

Prescriptive pricing is one of the most advanced pricing techniques, which derives the optimal price strategy to maximize the future profit/revenue by carrying out a two-stage process, demand modeling and price optimization. Demand modeling tries to reveal price-demand laws by discovering causal relationships among demands, prices, and objective factors, which is the foundation of price optimization. Existing methods either use regression or causal learning for uncovering the price-demand relations, but suffer from pain points in either accuracy/efficiency or mixed data type processing, while all of these are actual requirements in practical pricing scenarios. This paper proposes a novel demand modeling technique for practical usage. Speaking concretely, we propose a new locally consistent information criterion named MIC, and derive MIC-based inference algorithms for an accurate recovery of causal structure on mixed factor space. Experiments on simulate/real datasets show the superiority of our new approach in both price-demand law recovery and demand forecasting, as well as show promising performance to support optimal pricing.

1 Introduction

Prescriptive pricing [Caro and Gallien, 2012; Ito and Fujimaki, 2016; 2017] is one of the most advanced pricing techniques, which derives the optimal price strategy to maximize future profit/revenue on the basis of demand forecasting. Prescriptive pricing generally carries out a two stage analysis, 1) demand modeling that reveals the exact price-demand laws by discovering the quantitative causal relationships among prices, sales, and objective factors from historical observations, and 2) pricing optimization that finds the optimal price strategy by resolving a mathematical optimization problem constructed from the discovered relations. Demand modeling plays a vital role in pricing. A correct detection of actual factors affecting sales and an unbiased estimation of their quantitative effects by demand modeling, act as the foundation of the second stage, and decide whether pricing optimization can choose key control factors for optimal pricing. (biased relation estimates leads to sub-optimal pricing solutions.)

Several issues are crucial and should be resolved in demand modeling for practical usage. Firstly, factors involved in pricing usually comprise *a mixture of discrete (e.g., season) and continuous variables (e.g., prices, sales)*, needing a technique that can process mixed data types jointly. Secondly, demand modeling requires an *accurate* recovery of the quantitative price-demand relations, which is the foundation to achieve optimal price strategy. Thirdly, the demand modeling process should be *efficient* so as to timely catch the insight from the fast-changing market, especially to support online pricing.

The causal models targeting at causal relation discovery is a nature selection for discovering price-demand laws, and have been attracting extensive attention recently. A causal model describes the demand as a causal effect of its own price, the prices of competing products, promotion, seasonality, and etc. But the widely used causal models are essentially regressions [Lee, 2011; Caro and Gallien, 2012; Ferreira *et al.*, 2015]. Regression for correlation analysis is not causal discovery in nature, and pseudo causes are always reported due to confounders, leading to a low accuracy of price-demand relation recovery. Speaking concretely, correlation analysis will produce a high miss-hunting ratio of actual causes and much biased estimations of causal effects.

Causal structure discovery is able to eliminate pseudo causes by telling causation from correlation. Structural equation models (SEM) and causal Bayesian networks are two main classes of causal models [Pearl, 2009]. Recently, a family of SEMs called Causal Additive Noise Models (ANM) [Shimizu *et al.*, 2011; Bühlmann *et al.*, 2014] have been widely applied in causal structure discovery of continuous variables. ANM also can deal with discrete variables [Peters *et al.*, 2011a], but the case involving a mix of continuous and discrete variables has received little attention. Causal Bayesian network gives a probabilistic interpretation of causal relations. Independence-based methods like PC [Spirtes *et al.*, 2000] and score-based methods like GES [Chickering, 2003] are able to identify the causal graph from the joint distribution up to Markov equivalence, under the *causal Markov condition* and *faithfulness condition* [Pearl, 2009]. There has been a long line of work on recovering a hybrid Bayesian network from a mixture of variables by discretizing continuous variables [Monti and Cooper, 1998; Dojer, 2016], or converting the condition-

[*] Equal contribution. Corresponding author: Liu Chunchen.

al distributions of all variables into the same type [Romero *et al.*, 2006]. Until recently, some work [Cui *et al.*, 2016; Sedgewick *et al.*, 2017] adopted conditional independence-based methods to infer causal structures from mixed data.

In this paper, we develop a mixed causal model to describe price-demand relationships, propose a novel information criterion and derive related inference algorithms, to tackle the general issues required by demand modeling for practical usage. Our contributions can be detailed as follows:

Mixed causal model: We develop a mixed causal model for a mixture of discrete and continuous variables, and prove its identifiability in the bivariate cases.

Mixed information criterion: For score-based causal discovery, the key is an information criterion used for evaluating how mixed data fit a causal graph, which should be factorizable and tractable to compute. We derive a locally consistent criterion named Mixed Information Criterion (MIC) to tackle this problem.

Factorized causal inference: Based on MIC, we derive an inference framework for efficient structure discovery. Speaking concretely, with MIC, we construct a factorized optimization problem, on which an explorative search can be adopted to recover the causal structure. To accelerate the search process further, we propose ancestor-based search space cutting for speeding up without losing accuracy.

Evaluation on simuation/real retail datasets: We combine the proposed demand modeling technique with a latest pricing optimization method for test on both simulation and real datasets. Experiment results show that comparing with existing causal methods, our technique shows superiority in price-demand law recovery, supports demand prediction with higher accuracy, and helps to achieve higher gross profit.

2 Mixed Causal Modeling

Suppose we observe all factors relating to pricing, and have historical data $\mathbf{X} = [X_1, \cdots, X_D]$, where each random variable X_i can be binary ($X_i \in \{0, 1\}^N$), categorical ($X_i \in \{1, ..., T\}^N$), or continuous ($X_i \in \mathbb{R}^N$). Since a categorical variable with T class can be translated to $(T-1)$ binary variables equally by standard practice, $\mathbf{X}$ is simplified as a mixture of binary and continuous variables. Under the causal markov and faithfulness assumptions, we model the relationships among the D observed variables as a probabilistic graphical model defined over a directed acyclic graph (DAG) G. Each random variable X_i corresponds to a node in G, and there is an edge linking from X_i to X_j if X_i is a direct cause of X_j. The probability graph model is in the form,

$$\begin{aligned} & p(X_1, \cdots, X_D) \\ & = \Pi_{i=1}^{D} p_b(X_i|Pa(X_i))^{z_i} p_c(X_i|Pa(X_i))^{(1-z_i)}, \end{aligned} \tag{1}$$

where $Pa(X_i)$ is the parent set of X_i in G. $p_b(\bullet)$ and $p_c(\bullet)$ 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.

We assume the relations between a continuous variable and its parents are linear as (2) shows. Actually, the relations in real scenarios are generally non-linear, however, defining non-linear relations in demand modeling will raise great difficulty in the next pricing optimization stage, that is non-linear formulas make optimization problem hard to build and resolve or even lead the problem to be unsolvable. Alternatively, we represent non-linear by combining linear model and non-linear transformation of variables, which can avoid the above dilemma.

$$X_i = \boldsymbol{\beta}_i^T X_{-i} + \epsilon_i, \epsilon_i \sim Laplace(0, b_i), \tag{2}$$

β_i quantitatively describes the relations between X_i and all the other variables X_{-i} where $\beta_{ij} = 0$ for $X_j \notin Pa(X_i)$. Here we use a symmetric super-Gaussian distribution, *Laplace* distribution, which will produce a least absolute deviation score that is robust to outliers, and has been reported work well in *non-Gaussian* causality estimation methods including LiNGAM [Hyvärinen and Smith, 2013], S-LIM [Henao and Winther, 2011], and etc. We assume the relations between a binary variable and its parents as follows,

$$X_i = \begin{cases} 1 & \boldsymbol{\beta}_i^T X_{-i} + \epsilon_i > 0 \\ 0 & \text{otherwise} \end{cases}, \epsilon_i \sim Logistic(0, 1). \tag{3}$$

By introducing (2) and (3) to (1), we obtain the joint probability distribution,

$$\begin{aligned} & p(X_1, \cdots, X_D) \\ & = \Pi_{i=1}^{D} \Pi_{n=1}^{N} e^{\boldsymbol{\beta}_i^T X_{-i,n} x_{in} z_i} (1 + e^{\boldsymbol{\beta}_i^T X_{-i,n}})^{-z_i(1-x_{in})} \\ & \quad \Pi_{n=1}^{N} b_i^{z_i - 1} e^{-\frac{|x_{in} - \boldsymbol{\beta}_i^T X_{-i,n}| \cdot (1-z_i)}{b_i}}, \end{aligned} \tag{4}$$

where x_{in} is the n-th element of X_i.

2.1 Identifiability of the Mixed Causal Model

Here we give the definition of bivariate identifiability and prove the above mixed causal model is bivariate identifiable.

Definition 1 (Bivariate Identifiability) *[Peters* et al.*, 2011b] Let $\mathcal{F} = \{f | f : \mathbb{R}^2 \to \mathbb{R}\}$ be a set of functions, $\mathcal{P}_{\mathbb{U}} = \{\mathcal{P}_{\mathbb{R}}, \mathcal{P}_{\{0,1\}}\}$ denotes the set of probabilistic distributions for continuous/binary random variables, we call a set $\mathcal{B} \subseteq \mathcal{F} \times \mathcal{P}_{\mathbb{U}} \times \mathcal{P}_{\mathbb{U}}$ containing functions $f \in \mathcal{F}$ and distributions of inputs $\mathbf{X}$ and noise ε bivariate identifiable in $\mathcal{F}$ if*

$$\begin{aligned} & (f, P_X, P_{\varepsilon_Y}) \in \mathcal{B} \; and \; Y = f(X, \varepsilon_Y), X \perp\!\!\!\perp \varepsilon_Y \\ & \Rightarrow \nexists (g, P_Y, P_{\varepsilon_X}) \in \mathcal{B} \; and \; X = g(Y, \varepsilon_X), Y \perp\!\!\!\perp \varepsilon_X \end{aligned} \tag{5}$$

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

Theorem 1 *For the mixed causal model, $\mathcal{F} = \{f_c, f_b | f_c(x, \varepsilon) = \beta x + \varepsilon, f_b(x, \varepsilon) = \begin{cases} 1 & \beta x + \varepsilon > 0 \\ 0 & otherwise \end{cases}\}$, $\mathcal{P}_{\mathbb{U}}$ is a set of non-Gaussian distributions with Laplace distribution for noise on continuous variables and Logistic distributions for noise on binary variables. The model is bivariate identifiability.*

Here we provide an intuition rather than a formal proof which can be found in Appendix A. Briefly speaking, i) for the case that both variables are continuous, the model degenerates to a kind of linear ANM model of which the identifiability has been well proved by existing work [Shimizu *et al.*, 2006]; ii) for the case that both variables are binary, with a sufficient condition that the random variables do not share the same marginal distribution, the model is identifiable; iii) for the case of a mixed of binary and continuous variables, based on the differences in conditional distributions, the model is identifiable.

3 Causal Inference

3.1 Mixed Information Criterion (MIC)

In the score-based causal inference approach, a scoring function is defined over the space of DAG, and one searches this space for a structure that optimize the scoring function. While the traditional approaches estimate the graph structure G and parameters $\{\beta_1, \ldots, \beta_D\}$ jointly using scoring functions like the Bayesian Dirichlet likelihood-equivalent (BDe) [Chickering, 2003], we propose a mixed information criterion (MIC), to assess the fit of a DAG on a mixture of binary and continuous variables, as follows,

$$\begin{aligned} MIC(G) &= \sum_{i=1}^{D} MIC(X_i, Pa(X_i)) \\ &= \sum_{i=1}^{D} \left(\frac{1}{w_i} LL(X_i|Pa(X_i)) + Pen(X_i, Pa(X_i)) \right), \end{aligned} \quad (6)$$

where $LL(X_i|Pa(X_i))$ is the negative log-likelihood of data, and $Pen(X_i, Pa(X_i)$ is the ℓ_0 penalty. We introduce a scale parameter $\{w_i\}_{i=1}^{D}$ to make the negative log-likelihood of different variables comparable in terms of magnitude. The magnitude of the negative log-likelihood changes a lot from one variable to another, especially when the observations are a mixture of binary and continuous variables. Such magnitude variation may comes from the differences in the value of variables which are measured on different scales. Normalization is one common way to adjust different variables to an aligned scale, but it might bring undesirable changes into the probability distribution of adjusted variables, which will lead to an accuracy loss of causal structure estimate. Our introduced scale parameter $\{w_i\}_{i=1}^{D}$ is an alternative way for scale aligning while avoiding changing the data distribution. The magnitude variation may also comes from the estimation errors of model parameters, which will further bias the estimate of causal structure. Using a relative loss as MIC does is more robust to such magnitude variation than using an absolute loss if the scale parameter is proper designed. Here, we fix the value of w_i to an underestimate of $LL(X_i|Pa(X_i))$, that is,

$$Pa_c(X_i) = \arg\min_{X' \subseteq X_{-i}} LL(X_i|X'), \quad (7)$$

$$w_i = LL(X_i|Pa_c(X_i)), \quad (8)$$

$Pa_c(X_i)$ is the optimal potential parent set of variable X_i, and $Pa(X_i) \subseteq Pa_c(X_i)$. The relative loss $LL(X_i|Pa(X_i))/w_i$ measures the ratio of the absolute loss to an optimistic estimate, and is robust to data scale variations and estimation errors as validated by the experiments.

We give the definition of local consistency and prove that MIC score is locally consistent. Local consistency, which was proposed in [Chickering, 2003], means that optimizing the model selection criterion leads to select a graph that can represent the data generating distribution, and the graph contains no edges that are redundant for representing the data generating distribution [Schulte and Gholami, 2017].

Definition 2 (Local Consistency) *[Chickering, 2003] Let* $\mathbf{X}$ *be a set of data consisting of* N *records that are iid samples from some distribution* $p(\cdot)$*. Let* $\mathcal{G}$ *be any DAG, and let* $\mathcal{G}'$ *be the DAG that results from adding the edge* $X_i \rightarrow X_j$*. A scoring criterion* $S(\mathcal{G}, \mathbf{X})$ *is locally consistent if the following two properties hold:*

1. *If* $X_j \not\perp\!\!\!\perp_p X_i \mid Pa_j^{\mathcal{G}}$*, then* $S(\mathcal{G}', \mathbf{X}) > S(\mathcal{G}, \mathbf{X})$
2. *If* $X_j \perp\!\!\!\perp_p X_i \mid Pa_j^{\mathcal{G}}$*, then* $S(\mathcal{G}', \mathbf{X}) < S(\mathcal{G}, \mathbf{X})$

where, $Pa_j^{\mathcal{G}}$ is shorten for $Pa(X_j)$ in the DAG $\mathcal{G}$.

Theorem 2 *MIC score is locally consistent.*

The formal proof can be found in Appendix B.

3.2 MIC-Based Causal Inference

With the new defined MIC used for evaluating the goodness of causal structures, we construct an optimization formulation to model the problem of causal structure discovery from mixed data. Speaking concretely, by instantiating the log-likelihood in (6) with the joint probability distribution in (4), we obtain a target for optimization, and a DAG constraint and ℓ_0 constraints are added to achieve a sparse network.

$$\begin{aligned} &\min_{\beta_1, \cdots, \beta_D} \sum_{i=1}^{D} MIC(\beta_i, X_i, X_{-i}) \\ &\text{s.t.} \quad G_{\{\beta_1, \cdots, \beta_D\}} \in DAG, \| \beta_i \|_0 \leq k, \ i \in \{1, \cdots, D\}, \end{aligned}$$

$$MIC(\beta_i, X_i, X_{-i}) = \frac{LL(\beta_i, X_i, X_{-i})}{\min_{\beta_i, S(\beta_i) \subseteq X_{-i}} LL(\beta_i, X_i, X_{-i})},$$

$$\begin{aligned} LL(\beta_i, X_i, X_{-i}) &= (1 - z_i) \sum_{n=1}^{N} \left(\log b_i + \frac{|x_{in} - \beta_i^T X_{-i,n}|}{b_i} \right) \\ &+ z_i \sum_{n=1}^{N} \left((1 - x_{in}) \log(1 + e^{\beta_i^T X_{-i,n}}) - x_{in} \beta_i^T X_{-i,n} \right). \end{aligned} \quad (9)$$

$S(\beta_j)$ denotes the support set of β_j. When resolving formula (9), the challenge is how to minimize an objective while enforcing the implied graph structure should contain no directed cycles. Existing researches [Koller and Friedman, 2009] show that the DAG-constraint structure learning problem can be cast as that of learning an optimal ordering of variables. Once the variable order is fixed, the constraint of no directed cycles can be enforced by constraining the parents of a variable to be a subset of variables ordering precede it. Binding with our case, we translate the problem in formula (9) to a dynamic decision problem, of which the resolving will simultaneously produce the optimal variable ordering and a DAG-satisfied sparse causal structure.

$$\begin{aligned} orderSc(\boldsymbol{U}) = \min_{X_j \in \boldsymbol{U}} \Big\{ & orderSc(\boldsymbol{U} \setminus X_j) \\ & + nodeSc\left(X_j | \underline{(\boldsymbol{X} \setminus \boldsymbol{U}) \cap Pa_c(X_j)}\right) \Big\}, \end{aligned} \qquad (10)$$

$$\begin{aligned} & nodeSc\left(X_j | (\boldsymbol{X} \setminus \boldsymbol{U}) \cap Pa_c(X_j)\right) \\ = & \min_{\beta_j, S(\beta_j) \subseteq (\boldsymbol{X} \setminus \boldsymbol{U}) \cap Pa_c(X_j)} MIC(\beta_j, X_j, X_{-j}) \\ & \text{s.t. } \| \beta_j \|_0 \leq k, \end{aligned} \qquad (11)$$

$\boldsymbol{U}$ ($\boldsymbol{U} \subseteq \boldsymbol{X}$) is a set of variables of which the order *have not* been identified, $orderSc(\bullet)$ is the score of the optimal ordering of $\bullet$, $nodeSc(\bullet|\bar{\bullet})$ is the optimal score of $\bullet$ under the choice of its parents from $\bar{\bullet}$, $Pa_c(X_j)$ is the optimal potential parent set for variable X_j, whose computation is described as formula (7). Take $(\boldsymbol{X} \setminus \boldsymbol{U}) \cap Pa_c(X_j)$ as candidate parents for computing $nodeSc$ for X_j (formula (10)) will not reduce the causal discovery accuracy compared with using $\boldsymbol{X} \setminus \boldsymbol{U}$, but will save computation cost greatly (experiments show this further). We resolve the ℓ_0 regression problems (11) by applying the forward backward greedy (FoBa) algorithm [Zhang, 2011] since it offers the tightest upper bounds of feature selection error, estimation error, and objective error. Although the upper bounds for the original FoBa algorithm have only been derived for least square regression, we can achieve the same bounds for our problem by slightly modifying the proofs of [Zhang, 2011]. Please refer to Appendix A for how we adapt FoBa to our problem.

A* FoBa for Optimization

For the dynamic decision problem like formula (10), some previous work [Xiang and Kim, 2013] solves it by finding a shortest path in the order search space. The *start* state of this order search space is an empty variable set, the *goal* state is the complete variable set, and any other state represents a subset of the complete variable set. In this order search space, an arc is always linking from one state to another with one more variable added. Each arc has been assigned a cost, and a path from *start* to *goal* with the lowest total cost is called the "shortest" path. The order according to which variables are added on the "shortest" path is the optimal variable order. We borrow the idea and derive our own method A* FoBa for the MIC-based causal inference. It is called A* FoBa because we resolve the optimization problem in (10) by A* search, and solve the sparse regression problems in (11) using FoBa algorithm. Start from the *start* state in the order space, A* search explores a shortest path, by greedy selecting the most promising succeeding state. A state $\boldsymbol{Q}$ is called most promising if its estimated cost, $f(\boldsymbol{Q}) = g(\boldsymbol{Q}) + h(\boldsymbol{Q})$, is the lowest. $g(\boldsymbol{Q})$ summarizes the cost of the arcs on the path from the *start* to the $\boldsymbol{Q}$ state, which denotes the cost incurred so far. $h(\boldsymbol{Q})$ is the estimated future cost to be incurred from $\boldsymbol{Q}$ to the goal. Based on formula (10), we derive $g(\boldsymbol{Q})$ and $h(\boldsymbol{Q})$ as following,

$$g(\boldsymbol{Q}) = \sum_{X \in \boldsymbol{Q}} nodeSc(X | \Pi^{\boldsymbol{Q}}_{\prec X} \cap Pa_c(X)), \qquad (12)$$

$$h(\boldsymbol{Q}) = \sum_{X \in \boldsymbol{X} \setminus \boldsymbol{Q}} nodeSc(X | \boldsymbol{X} \setminus X). \qquad (13)$$

$\Pi^{\boldsymbol{Q}}$ denotes an ordering of the nodes in $\boldsymbol{Q}$ and $\Pi^{\boldsymbol{Q}}_{\prec X}$ denotes the set of variables in $\boldsymbol{Q}$ that precede X in $\Pi^{\boldsymbol{Q}}$. A* search requires the h function to be *admissible*, meaning that $h(\boldsymbol{Q})$ is always an underestimate of the true cost to reach the goal. Also, the h function should be *consistent*, meaning it should satisfy $h(\boldsymbol{Q}) \leq c(\boldsymbol{Q}, \boldsymbol{Q}') + h(\boldsymbol{Q}')$, where $\boldsymbol{Q}' = \boldsymbol{Q} \cup \{X_j\}$ is a succeeding state and $c(\boldsymbol{Q}, \boldsymbol{Q}') = nodeSc(X_j | \boldsymbol{Q} \cap Pa_c(X_j))$ is the cost of moving from $\boldsymbol{Q}$ to $\boldsymbol{Q}'$. We prove that, based on score definitions((11), (13)) and the working process of the FoBa algorithm, $h(\boldsymbol{Q})$ in (13) is *admissible* and *consistent*.

Theorem 3 *The h function in equation* (13) *is admissible.*

Proof: The true cost from the current state to the goal is $T(\boldsymbol{Q}) = \sum_{X \in \boldsymbol{X} \setminus \boldsymbol{Q}} nodeSc(X | \Pi_{\prec X} \cap Pa_c(X))$. The parent set selected by $nodeSc(X | \boldsymbol{X} \setminus X)$ in h is $Pa_c(X)$ based on the definition in (7), while the parent set selected by $nodeSc(X | \Pi_{\prec X} \cap Pa_c(X))$ is a subset of $Pa_c(X)$ and denoted by $Pa'_c(X)$. Considering the feature selection process of FoBa, a feature $r \in Pa_c(X) \setminus Pa'_c(X)$ will be selected only if $nodeSc$ gets smaller. Therefore, $h(\boldsymbol{Q}) \leq T(\boldsymbol{Q})$ and the h function is *admissiable*. ■

Theorem 4 *The h function in equation* (13) *is consistent.*

Based on Theorem 1, Theorem 2 can be easily proved, and we skip it because of the space constraint.

3.3 Ancestor-Based Search Space Cutting

We utilize the ancestor relations to further cut down the order search space for inference accelerating.

Theorem 5 *Ancestor relations can be used to prune the search space without losing optimality.*

Proof: Given $\boldsymbol{Q}$ the set of variables assigned in the current state, and $\boldsymbol{Q} \cap \{X_i, X_j\} = \emptyset$, there are two paths to achieve the future state with variables $\boldsymbol{Q} \cup \{X_i, X_j\}$. $Path_1$ is $\boldsymbol{Q} \to \boldsymbol{Q} \cup \{X_i\} \to \boldsymbol{Q} \cup \{X_i, X_j\}$, and $Path_2$ is $\boldsymbol{Q} \to \boldsymbol{Q} \cup \{X_j\} \to \boldsymbol{Q} \cup \{X_i, X_j\}$. If the ancestor relation between two variables, e.g. $X_i \prec X_j$, is known, then we can eliminate $Path_2$ because it is always suboptimal compared with $Path_1$ which is proved as follows,

$$\begin{aligned} & cost(Path_2) - cost(Path_1) \\ = & nodeSc(X_j \mid \boldsymbol{Q} \cap Pa_c(X_j)) - nodeSc(X_i \mid \boldsymbol{Q} \cap Pa_c(X_i)) \\ & + nodeSc(X_i \mid \boldsymbol{Q} \cup \{X_j\} \cap Pa_c(X_i)) \\ & - nodeSc(X_j \mid \boldsymbol{Q} \cup \{X_i\} \cap Pa_c(X_j)) \\ = & nodeSc(X_j \mid \boldsymbol{Q} \cap Pa_c(X_j)) \\ & - nodeSc(X_j \mid \boldsymbol{Q} \cup \{X_i\} \cap Pa_c(X_j)) \geq 0, \end{aligned} \qquad (14)$$

in which $nodeSc(X_i \mid \boldsymbol{Q} \cup \{X_j\} \cap Pa_c(X_i))$=$nodeSc(X_i \mid \boldsymbol{Q} \cap Pa_c(X_i))$ due to the ancestor constraint $X_i \prec X_j$. ■

Algorithm 1 A* FoBa with ancestor-based space cutting

```
Input: Data $\boldsymbol{X}$, the number of variables $D$
Output: Optimal structure $\boldsymbol{B} = \{\beta_d\}_{d=1}^{D}$
Initialize: empty queue open, empty set close
 1: For each variable $X_j \in \boldsymbol{X}$, get its optimal potential parent set
    $Pa_c(X_j)$ by solving (7) using FoBa
 2: Construct an adjacent matrix from $\{Pa_c(X_j)\}_{j=1}^{D}$
 3: Extract $m$ SCCs of variables $scc_1 \prec scc_2 \prec \cdots \prec scc_m$ from
    the adjacent matrix using Tarjan's algorithm [Tarjan, 1972]
 4: Project SCCs onto a series ordering constraints $\boldsymbol{C} = \{X_i \prec X_j \mid X_i \in scc_k, X_j \in scc_l, scc_k \prec scc_l\}$, $\forall i, j \in$
    $\{1, \cdots, D\}$,$\forall k, l \in \{1, \cdots, m\}$
 5: open.insert($\boldsymbol{Q} = \emptyset, f(\emptyset) = h(\emptyset), g(\emptyset) = 0, \boldsymbol{B}_Q = \emptyset$)
 6: while true do
 7:    $\boldsymbol{Q}, f(\boldsymbol{Q}), g(\boldsymbol{Q}), \boldsymbol{B}_Q \leftarrow$ open.pop()
 8:    if $h(\boldsymbol{Q}) = 0$ then  return $\boldsymbol{B}_Q$  end if
 9:    for each $X_j \in \boldsymbol{X} \setminus \boldsymbol{Q}$ do
10:       $\boldsymbol{Q}' \leftarrow \boldsymbol{Q} \cup \{X_j\}$
11:       if $\boldsymbol{Q}' \notin close$ and not $voilate(\boldsymbol{Q}', \boldsymbol{C})$ then
12:          Compute $nodeSc(X_j \mid \boldsymbol{Q} \cap Pa_c(X_j))$ by (11)
13:          $g(\boldsymbol{Q}') \leftarrow g(\boldsymbol{Q}) + nodeSc(X_j \mid \boldsymbol{Q} \cap Pa_c(X_j))$
14:          $h(\boldsymbol{Q}') \leftarrow h(\boldsymbol{Q}) - nodeSc(X_j \mid \boldsymbol{X} \setminus X_j))$
15:          $f(\boldsymbol{Q}') \leftarrow g(\boldsymbol{Q}') + h(\boldsymbol{Q}'), \boldsymbol{B}_{Q'} \leftarrow \boldsymbol{B}_Q \cup \{\beta_j\}$
16:          open.insert($\boldsymbol{Q}', f(\boldsymbol{Q}'), g(\boldsymbol{Q}'), \boldsymbol{B}_{Q'}$)
17:          $close \leftarrow close \cup \{\boldsymbol{Q}'\}$
18:       end if
19:    end for
20: end while
```

We first learn the ancestor relations among all the variables, and then project these ancestor relations onto a series of topological ordering constraints, which can be easily integrated into the A* FoBa framework for search space cutting. The whole algorithm is summarized in Algorithm 1. The function $voilate(\boldsymbol{Q}', \boldsymbol{C})$ in line 11 checks whether the current variable ordering violates the topological ordering constraints. The time complexity of our MIC-based causal inference algorithm is $O(m \max_i 2^{|scc_i|} D^2 N)$, where m is the number of strongly connected components (SCCs) [Tarjan, 1972], $|scc_i|$ is the number of variables in the SCC scc_i, D denotes the total number of variables, N is the sample number.

4 Simulated Study

4.1 Synthetic Data

We simulated prices and sales of 40 kinds of beer, and 9 objective factors (i.e., *temperature* (continuous), *seasonality* (4, binary), *vacation* (binary), *promotion* (binary) and *sunny/rainy* (2, binary)). We used p_m^t and q_m^t to denote the price and sales of the m-th product in the t-th day, and denoted the n-th objective factor in the t-th day as g_n^t. To mimic the nonlinear relations between prices and sales, we introduced univariate transformation on prices that are $f_1(p_m^t) = 1/p_m^t$, $f_2(p_m^t) = p_m^t$, and $f_3(p_m^t) = \ln(p_m^t)$. Prices and sales are simulated by (15) and (16). The coefficient $\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$ were randomly generated from reasonable ranges. For noisy terms $\boldsymbol{\epsilon}$ or $\boldsymbol{e}$, we simulated exponential distribution and mixture Gaussian distribution.

$$q_m^t = \alpha_m^* + \sum_{m'=1}^{40} \sum_{i=1}^{3} \beta_{mm'i}^* f_i(p_{m'}^t) + \sum_{n=1}^{9} \gamma_{mn}^* g_n^t + \epsilon_m^t, \quad (15)$$

$$p_m^t = a_m^* + \sum_{m' \neq m} \sum_{i=1}^{3} b_{mm'i}^* f_i(p_{m'}^t) + \sum_{n=1}^{9} c_{mn}^* g_n^t + e_m^t. \quad (16)$$

We simulated 16 experiment settings differed in noise types, sparsity (simple/complex structure), train sample scale (1/3 years), and test sample scale (7/30 days). $\mathbf{b}^* = \mathbf{0}$ holds in the *simple structure*, while $\mathbf{b}^*$ is not zero in *complex structure*. Under each setting, we generated 10 datasets while each contained 169 random variables.

4.2 Benchmarks

We compared our method with 7 other methods, 1) the $l0$ sparse regression [Liu *et al.*, 2014] as a representative of regression-based demand modeling, 2) stable PC [Colombo and Maathuis, 2014] that is an improved version of the classical constraint-based method, PC, 3) causalMGM [Sedgewick *et al.*, 2017], which is a latest constraint-based method for mixed data processing, 4) GES [Chickering, 2003] that is a classical score-based method, 5) A* Lasso [Xiang and Kim, 2013] whose shortest path finding framework we refer to, 6) Direct LiNGAM [Shimizu *et al.*, 2011] that is designed for linear non-Gaussian data, 7) CAM [Bühlmann *et al.*, 2014] that is a nonlinear-Gaussian SEM. Besides implementing the $l0$ sparse regression by ourselves, we used the implementations provided by authors and their default parameter settings for all the other benchmarks. The stable PC and causalMGM use the likelihood ration test (LRT) for independent test.

4.3 Evaluation Metrics

We consider the following evaluations:

Causal structure discovery (price-demand law discovery): Two metrics were used to measure the accuracy of structure discovery: 1) $Precision = \frac{TP}{TP+FP}$, denoting among all the discovered causal relations, how many actually hold, and 2) $Recall = \frac{TP}{TP+FN}$, denoting among all the actual causal relations, how many have been identified.

Demand forecasting: We used the learnt demand models to predict sales of 40 products on the test samples, and we used the averaged root-mean-square error (RMSE) to measure the prediction accuracy.

Pricing strategy: Based on a learnt demand model, we defined the optimization target, gross profit, as

$$l(\mathbf{p}) = \sum_{t=1}^{T} \sum_{m=1}^{M} (p_m - r_m) q_m^t(\mathbf{p}, \mathbf{g}^t), \quad (17)$$

where $\mathbf{p} = \{p_m\}_{m=1}^{M}$ are control variables to be optimized, $\mathbf{g} = \{g_n\}_{n=1}^{N}$ are external variables whose values are already known from test samples. $q_m^t(\bullet)$ is the function revealing price-demand relations. T is the number of test samples, r_m is the cost of the m-th product. With the gross profit function as objective and with some actual requirements as constraints (e.g., p_m is chosen from discrete values $\{p_{m1}, \ldots, p_{mJ}\}$.), the pricing optimization problem was formulated as binary quadratic programming (BQP), and solved by a semi-definite programming (SDP) relaxation [Ito and Fujimaki, 2017].

For evaluation, let $l^*(\bullet)$ denotes a gross profit function with true parameters $\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$, then the true optimal pricing strategy is $\mathbf{p}^* = \arg\max_{\mathbf{p}} l^*(\mathbf{p})$. Let $\hat{l}(\bullet)$ denotes a profit function on $\left\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}\right\}$ that are learnt by a demand model from train data, then a generated pricing strategy is $\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \hat{l}(\mathbf{p})$. We used $\frac{l^*(\hat{\mathbf{p}})}{l^*(\mathbf{p}^*)} \in [0, 1]$ to measure the goodness of a pricing strategy. ($l^*(\hat{\mathbf{p}})$ means the actual profit if running strategy $\hat{\mathbf{p}}$ in real environments, while $l^*(\mathbf{p}^*)$ means the ideal profit if using optimal strategy $\mathbf{p}^*$.)

4.4 Experimental Results

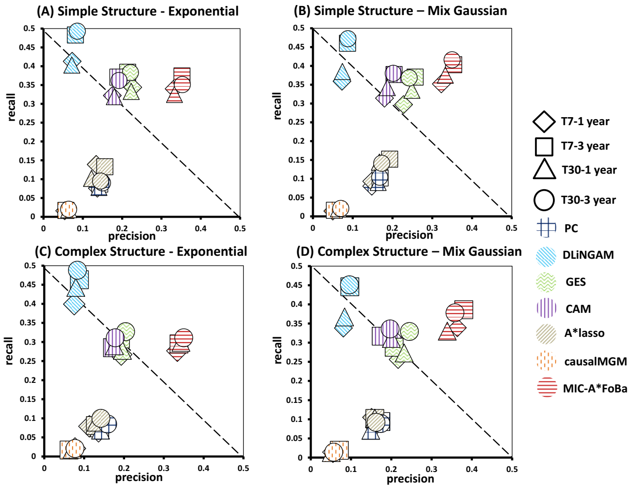

Figure 1: Comparisons of structure discovery.

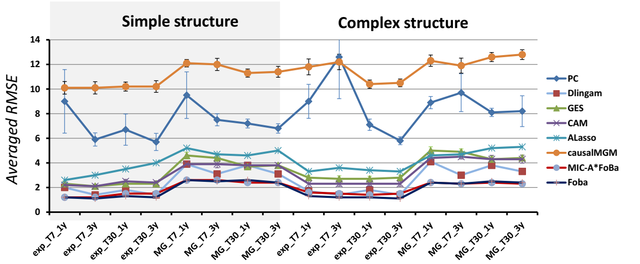

Figure 2: Comparisons of prediction accuracy. The horizontal axis indicates 16 exp. settings.

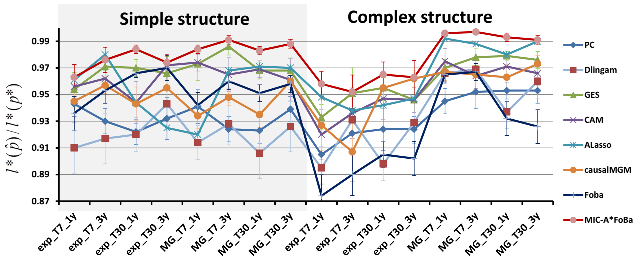

Figure 3: Comparisons of pricing accuracy.

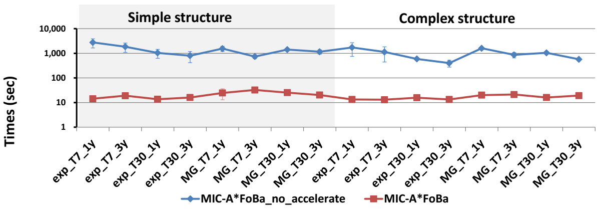

Figure 4: Comparisons of elapsed time between MIC-A*FoBa with and without the acceleration policies.

Figure 1 summarizes the accuracy comparisons for causal structure discovery. The four sub-figures tell us that, across all the datasets that vary in structure, noise distribution, train data scale, and test duration, our method (MIC-A*FoBa) outperformed all the other methods in *Precision* obviously and identically. Such results mean that the causal relations detected by our method are more reliable since a higher ratio of its identified relations are actually true. When checking the performance of *Recall*, we find that our method is better than (Figure 1 B/D) or comparable with (Figure 1 A/C) most of the benchmarks except the DLiNGAM. A high *Recall* while a low *Precision* like DLiNGAM means that it hunts a large amount of true and pseudo relations at the same time, which is bad news especially for the pricing optimization, because these pseudo causes will bias the causal effects of true causes which will further lead to bad pricing strategy.

Figure 2 compares the demand forecasting accuracy. Results show that MIC-A*FoBa defeats all the other causal methods across all the datasets and achieved the smallest prediction error. Comparing MIC-A*FoBa with FoBa that is for prediction, they show comparable accuracy, which further confirms that the causal relations and their related causal effects learnt by MIC-A*FoBa is close to the groundtruth.

Figure 3 compares the goodness of pricing strategies. MIC-A*FoBa shows the best performance across all datasets, meaning that it helps to produce pricing strategies that bring the highest gross profits. For the $l0$ regression, it worked relative well on datasets with simple structures, but dropped sharply on complex structures. One possible explanation is that, the regression failed to find real causes for demand variation due to price interactions in complex structure, and such pseudo causes lead to produce sub-optimal pricing strategy.

In A*FoBa, we adopted two policies for accelerating the inference. One is by the ancestor-based search space cutting strategy, and the other is by adding $Pa_c(\bullet)$ to constraint the candidate parent space as the underlined part in formula (9) shows. We compare two versions of MIC-A*FoBa, one is with the above strategies adopted while the other does not. The comparison showed in Figure 4 confirms the usefulness of the policies for inference accelerating.

5 Real World Retail Data

5.1 Data and Experimental Settings

We applied our method to real retail data from a supermarket in Tokyo.[1] We used daily prices, univariate transformations

[1] The data is provided by KSP-SP Co.,LTD.

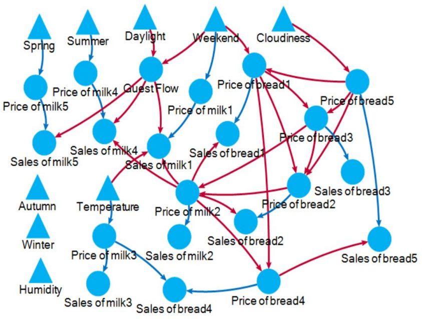

Figure 5: Causal structure on retail data. Red/blue lines indicate the positive/negative effect.

on prices and sales of 10 products (5 kinds of milk and 5 kinds of bread)[2] in two years from 2012/01 to 2013/12. In addition, objective factors during the same period were used, like the *daily guest flow volume* (continuous), *seasonality* (4, binary), *month* (12, binary), *weekend* (binary), *day of the week* (7, binary), *weather* (binary), and *temperature* (continuous). In total, we used 69 randomly variables for analysis.

We simulated the whole procedure of prescriptive pricing on this real data. For demand modeling, we learnt the sales formula of each product from the data of the year 2012. As for the price optimization, we constructed the optimization target, $\hat{l}(\mathbf{p})$, using (17) setting the cost to 0 for simplicity, and then obtained a pricing strategy $\hat{\mathbf{p}}$ by maximizing the revenue. To evaluate the pricing strategy, we built an independent validation environment $\hat{l}'(\bullet)$ by estimating new sale formulas on data from 2013/01 to 2013/09. Compared to $\hat{l}(\bullet)$, $\hat{l}'(\bullet)$ is an environment that is more similar to the real one in the validation stage (2013/10). We ran the pricing strategy $\hat{\mathbf{p}}$ on data from 2013/10, and achieved $\hat{l}'(\hat{\mathbf{p}})$ as the validation revenue explained as running $\hat{\mathbf{p}}$ in the validation sale environment.

5.2 Causal-Structure-Based Insight

A subgraph of the causal graph learnt by our method was shown in Figure 5 to make a concise presentation. It illustrates some laws of market, such as the price elasticity of demand (e.g. *price of bread1* has a negative effect on *sales of bread1*), the cross-price elasticity of demand (e.g. *price of milk2* has a positive effect on *sales of milk1*). It also reveals some relations that are in accord with common sense, such as the *guest flow* has a positive effect on sales, and the *weekend* and *daylight* have a positive effect on *guest flow*.

5.3 Improvement of Profit

Figure 6 **a** $\sim$ **c** compare the predicted sales of different methods with the true sales, we find the predicted sales curves of Dlingam, Alasso, PC and CAM are much biased from the true ones, indicating they discovered pseudo causes and/or biased causal effects. Figure 6 **d** compares the validation revenue by different methods with the actual revenue computed from historical data (the above 4 methods are excluded because they performed not good at pricing and for a clear visualization.).

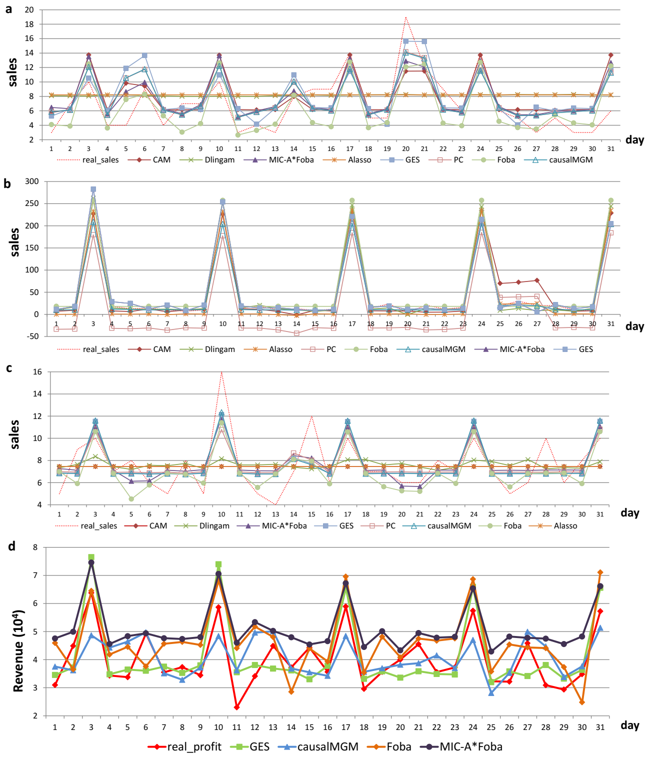

Figure 6: Comparison of sale forecasting and pricing strategy on 2013/10. **a**$\sim$ **c**. Predicted sales of 3 randomly chosen products by different methods. **d**. Validation revenues by different methods.

The sum of the validation revenue by GES, causalMGM, Foba, and MIC-A*Foba were 1.27×10^7, 1.26×10^7, 1.44×10^7, and 1.58×10^7 respectively. The revenue obtained by our method is the highest. When comparing the curve of MIC-A*Foba with actual one in detail, we find that the improvements of revenue in low-actual-revenue days is more significant than that in high-actual-revenue days. It suggests that our pricing strategy is more effective when the business is tepid.

6 Conclusion

We proposed a mixed information criterion MIC and derived an MIC-based causal inference method for accurate causal discovery on mixed factor space. Experiments showed our demand modeling method outperformed other competitors in terms of causal recovery, prediction and pricing strategy.

A Proof of Bivariate Identifiability

We respectively prove the bivariate identifiability of mixed causal model in the binary-binary case, continuous-binary case, continuous-continuous case.

Theorem 6 *Given a pair of binary variables (X, Y) all of which take values in $\{0, 1\}$, under the mild assumption that X and Y do not share the same marginal distribution, the mixed causal model is bivariate identifiable.*

Proof: If the mixed causal model is not bivariate identifiable, under the definition of the causal model and the defini-

[2] We used a few products to make causal graph visible and clear.

tion of bivariate identifiability, there exist two models,

$$\mathcal{M}_1 = \{Y = f_b(X, \varepsilon_Y) = \begin{cases} 1 & \beta_1 X + \varepsilon_Y > 0 \\ 0 & \text{otherwise} \end{cases},$$
$$P_X(X = 1) = k_1, P_{\varepsilon_Y} = logistic(0, 1)\}$$

$$\mathcal{M}_2 = \{X = f_b(Y, \varepsilon_X) = \begin{cases} 1 & \beta_2 Y + \varepsilon_X > 0 \\ 0 & \text{otherwise} \end{cases},$$
$$P_Y(Y = 1) = k_2, P_{\varepsilon_X} = logistic(0, 1)\}$$

of which the joint distributions are the same. Let us consider their joint distributions respectively, that are,

$$\begin{aligned} &P_{\mathcal{M}_1}(X, Y) = P_X(X)P_{\varepsilon_Y}(Y|X) \\ &= k_1^X(1 - k_1)^{1-X}(\frac{1}{1 + e^{-\beta_1 X}})^Y(1 - \frac{1}{1 + e^{-\beta_1 X}})^{1-Y}, \end{aligned}$$

$$\begin{aligned} &P_{\mathcal{M}_2}(X, Y) = P_Y(Y)P_{\varepsilon_X}(X|Y) \\ &= k_2^Y(1 - k_2)^{1-Y}(\frac{1}{1 + e^{-\beta_2 Y}})^X(1 - \frac{1}{1 + e^{-\beta_2 Y}})^{1-X}. \end{aligned}$$

If $P_{\mathcal{M}_1}(X, Y) = P_{\mathcal{M}_2}(X, Y)$ holds, it induces that $\frac{P_{\mathcal{M}_1}(X=0,Y=0)}{P_{\mathcal{M}_2}(X=0,Y=0)} = 1$, and it further induces $k_1 = k_2$, which conflicts with the assumption that the marginal distributions for X and Y should not be the same, so comes the theorem. ■

Theorem 7 *Given a pair of mixed variables* (X, Y) *that* X *takes value in* $\mathbb{R}$ *and* Y *takes value in* $\{0, 1\}$*, the mixed causal model is bivariate identifiable.*

Proof: If the mixed causal model is not bivariate identifiable, under the definition of the causal model and the definition of bivariate identifiability, there exist two models,

$$\begin{aligned} \mathcal{M}_1 =& \{Y = \begin{cases} 1 & \beta_1 X + \varepsilon_Y > 0 \\ 0 & \text{otherwise} \end{cases}, P_{\varepsilon_Y} = logistic(0, 1)\}, \\ \mathcal{M}_2 =& \{X = \beta_2 Y + \varepsilon_X, P_{\varepsilon_X} = Laplace(0, b)\}, \end{aligned}$$

of which the joint distributions are the same.

However, for model $\mathcal{M}_1$, we have $P(Y = 1|X = x) = P_{\varepsilon_Y}(\varepsilon_Y > -\beta_1 X)$ and $P_{\varepsilon_Y} = logistic(0, 1)$, then

$$\begin{aligned} &\lim_{x\to+\infty} P(Y = 1|X = x) = \begin{cases} 1, & \text{if } \beta_1 > 0 \\ 0, & \text{if } \beta_1 < 0 \end{cases}, \\ &\lim_{x\to-\infty} P(Y = 1|X = x) = \begin{cases} 0, & \text{if } \beta_1 > 0 \\ 1, & \text{if } \beta_1 < 0 \end{cases}. \end{aligned} \tag{18}$$

For model $\mathcal{M}_2$, from $P_{\varepsilon_X} = Laplace(0, b)$ and

$$\begin{aligned} &P(Y = 1|X = x) = \frac{P(Y = 1, X = x)}{P_X(X = x)} \\ &= \frac{P_Y(Y = 1)P_{\varepsilon_X}(X = x|Y = 1)}{P_Y(Y = 0)P_{\varepsilon_X}(X = x|Y = 0) + P_Y(Y = 1)P_{\varepsilon_X}(X = x|Y = 1)}, \end{aligned}$$

we can derive

$$\lim_{x\to\pm\infty} P(Y = 1|X = x) = P_Y(Y = 1), \tag{19}$$

which conflicts with that in (18), so comes the theorem. ■

Theorem 8 *Given a pair of continuous variables* (X, Y) *which all take values in* $\mathbb{R}$*, the mixed causal model is bivariate identifiable.*

Proof: For the situation where both variables are continuous, our model degenerate to a kind of linear additive noise model, whose identifiability has been well proved by [Shimizu *et al.*, 2006]. ■

B Proof of Local Consistency

Revisit the definition of decomposable MIC score in (6), $\left|Pa_j^{\mathcal{G}}\right|$ is the cardinality of $Pa_j^{\mathcal{G}}$,

$$MIC(\mathcal{G}, \mathbf{X}) = \sum_{j=1}^{D} \left(-\frac{1}{w_j} \log p(X_j \mid Pa_j^{\mathcal{G}}) + \lambda \left|Pa_j^{\mathcal{G}}\right| \right)$$

and BIC score is as follows,

$$BIC(\mathcal{G}, \mathbf{X}) = \sum_{j=1}^{D} \left(\log p(X_j \mid Pa_j^{\mathcal{G}}) - \frac{\log N}{2} \left|Pa_j^{\mathcal{G}}\right| \right)$$

Let $\mathbf{X}$ be a set of data consisting of N records that are iid samples from some distribution $p(\cdot)$. Let $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG that results from adding the edge $X_i \to X_j$,

$$\begin{aligned} &MIC(\mathcal{G}', \mathbf{X}) - MIC(\mathcal{G}, \mathbf{X}) \\ &= \frac{1}{w_j} \left(\log p(X_j \mid Pa_j^{\mathcal{G}}) - \log p(X_j \mid Pa_j^{\mathcal{G}} \cup \{X_i\}) \right) + \lambda. \end{aligned}$$

1. If $X_j \not\perp\!\!\!\perp_p X_i \mid Pa_j^{\mathcal{G}}$,

   $$\begin{aligned} &MIC(\mathcal{G}', \mathbf{X}) - MIC(\mathcal{G}, \mathbf{X}) \\ &= \frac{1}{w_j} \left(BIC(\mathcal{G}, \mathbf{X}) - BIC(\mathcal{G}', \mathbf{X}) \right) - \frac{\log N}{2w_j} + \lambda < 0. \end{aligned}$$

   As BIC score is locally consistent [Chickering, 2003], $BIC(\mathcal{G}, \mathbf{X}) - BIC(\mathcal{G}', \mathbf{X}) < 0$ follows the definition of local consistency, $w_j, \lambda \in \mathbb{R}^+$ are constant, $\log N \to +\infty$ as $N \to +\infty$.

2. If $X_j \perp\!\!\!\perp_p X_i \mid Pa_j^{\mathcal{G}}$,

   $$\begin{aligned} &p(X_j \mid Pa_j^{\mathcal{G}'}) = p(X_j \mid Pa_j^{\mathcal{G}} \cup \{X_i\}) = \frac{p(X_j, X_i \mid Pa_j^{\mathcal{G}})}{p(X_i \mid Pa_j^{\mathcal{G}})} \\ &= \frac{p(X_j \mid Pa_j^{\mathcal{G}})p(X_i \mid Pa_j^{\mathcal{G}})}{p(X_i \mid Pa_j^{\mathcal{G}})} = p(X_j \mid Pa_j^{\mathcal{G}}), \end{aligned}$$

   and thus,

   $$\begin{aligned} &MIC(\mathcal{G}', \mathbf{X}) - MIC(\mathcal{G}, \mathbf{X}) \\ &= \frac{1}{w_j} \left(\log p(X_j \mid Pa_j^{\mathcal{G}}) - \log p(X_j \mid Pa_j^{\mathcal{G}}) \right) + \lambda \\ &= \lambda > 0. \end{aligned}$$

Hence MIC score is locally consistent. ■

References

[Bühlmann *et al.*, 2014] Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[Caro and Gallien, 2012] Felipe Caro and Jérémie Gallien. Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422, 2012.

[Chickering, 2003] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(3):507–554, 2003.

[Colombo and Maathuis, 2014] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3921–3962, 2014.

[Cui *et al.*, 2016] Ruifei Cui, Perry Groot, and Tom Heskes. Copula pc algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer, 2016.

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

[Ferreira *et al.*, 2015] Kris Johnson Ferreira, Bin Hong-Alex. Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88, 2015.

[Henao and Winther, 2011] Ricardo Henao and Ole Winther. Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, 12(5):863–905, 2011.

[Hyvärinen and Smith, 2013] Aapo Hyvärinen and Stephen M. Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152, 2013.

[Ito and Fujimaki, 2016] Shinji Ito and Ryohei Fujimaki. Large-scale price optimization via network flow. *Advances in Neural Information Processing Systems*, pages 3855–3863, 2016.

[Ito and Fujimaki, 2017] Shinji Ito and Ryohei Fujimaki. Optimization beyond prediction: Prescriptive price optimization. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1833–1841, 2017.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. Probabilistic graph models: principles and techniques. *MIT press*, 2009.

[Lee, 2011] Seonah Lee. Study of demand models and price optimziation performance. *PhD theses, Georgia Institude of Technology*, 2011.

[Liu *et al.*, 2014] Ji Liu, Jieping Ye, and Ryohei Fujimaki. Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. *International Conference on Machine Learning*, pages 503–511, 2014.

[Monti and Cooper, 1998] Stefano Monti and Gregory F. Cooper. A multivariate discretization method for learning bayesian networks from mixed data. *Uncertainty in Artificial Intelligence*, pages 404–413, 1998.

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

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

[Peters *et al.*, 2011b] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Scholkopf. Identifiability of causal graphs using functional models. *Uncertainty in Artificial Intelligence*, pages 589–598, 2011.

[Romero *et al.*, 2006] Vanessa Romero, Rafael Rumí, and Antonio Salmerón. Learning hybrid bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42(1-2):54–68, 2006.

[Schulte and Gholami, 2017] Oliver Schulte and Sajjad Gholami. Locally consistent bayesian networks scores for multi-relational data. *International Joint Conference on Artificial Intelligence*, 2017.

[Sedgewick *et al.*, 2017] Andrew J. Sedgewick, Joseph D. Ramsey, Peter Spirtes, Clark Glymour, and Panayiotis V. Benos. Mixed graphical models for causal analysis of multi-modal variables. *arXiv preprint arXiv:1704.02621*, 2017.

[Shimizu *et al.*, 2006] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvarinen, and Antti J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[Shimizu *et al.*, 2011] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. Causation, prediction, and search. 2000.

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

[Xiang and Kim, 2013] Jing Xiang and Seyoung Kim. A* lasso for learning a sparse bayesian network structure for continuous variables. *Neural Information Processing Systems*, pages 2418–2426, 2013.

[Zhang, 2011] Tong Zhang. Sparse recovery with orthogonal matching pirsuit under rip. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.