

基于非纯测量场景的隐变量因果结构学习算法*

王馥弘¹, 陈正鸣², 夏业伟³, 乔杰¹, 郝志峰^{1,2}, 蔡瑞初¹

¹(广东工业大学 计算机学院, 广东 广州 510006)

²(汕头大学 理学院, 广东 汕头 515063)

³(复旦大学 计算与智能创新学院, 上海 200438)

通信作者: 蔡瑞初, E-mail: cairuichu@gmail.com



摘要: 隐变量间的因果结构学习, 核心在于从观测数据中挖掘隐变量彼此的因果关联, 是科学研究中揭示现象本质的一种关键技术. 现有方法普遍依赖“纯子”假设, 即隐变量对应的观测子代变量(测量变量)间不存在直接的因果连接边, 该假设在许多现实场景中往往无法成立, 从而导致现有方法可识别性受限. 针对该挑战, 考虑“非纯”测量场景下的隐变量识别性问题, 提出线性非高斯无环隐变量模型(linear non-Gaussian acyclic latent variable model, LiNGLM), 该模型允许观测变量间存在因果关联. 基于该模型, 提出一种隐变量结构学习(latent variable structure learning, LLSTIN)算法, 该算法基于变换独立噪声(transformed independent noise, TIN)条件及其建立的图准则, 首先通过因果聚类构建(build causal cluster, BCC)算法提取各隐变量对应的“单因子集”来识别隐变量的存在, 然后选取集合中的“根”观测变量作为隐变量的有效代理, 进而识别隐变量间的因果关系. 从理论上证明算法可正确提取单因子集并进一步识别隐变量因果结构, 仿真数据与真实数据实验结果进一步验证所提算法的正确性和有效性.

关键词: 因果结构学习; 线性非高斯; “非纯”测量场景; 变换独立噪声

中图法分类号: TP18

中文引用格式: 王馥弘, 陈正鸣, 夏业伟, 乔杰, 郝志峰, 蔡瑞初. 基于非纯测量场景的隐变量因果结构学习算法. 软件学报. <http://www.jos.org.cn/1000-9825/7633.htm>

英文引用格式: Wang FH, Chen ZM, Xia YW, Qiao J, Hao ZF, Cai RC. Causal Structure Learning Algorithm for Latent Variables in Impure Measurement Scenarios. Ruan Jian Xue Bao/Journal of Software (in Chinese). <http://www.jos.org.cn/1000-9825/7633.htm>

Causal Structure Learning Algorithm for Latent Variables in Impure Measurement Scenarios

WANG Fu-Hong¹, CHEN Zheng-Ming², XIA Ye-Wei³, QIAO Jie¹, HAO Zhi-Feng^{1,2}, CAI Rui-Chu¹

¹(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

²(College of Science, Shantou University, Shantou 515063, China)

³(College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200438, China)

Abstract: Learning the causal structure among latent variables is a key technique for revealing the underlying mechanisms of phenomena in scientific research, with its core objective being to infer causal relationships between latent variables from observational data. Existing methods generally rely on the “pure child” assumption, which assumes that no direct causal connections exist among the observed child variables (measurement variables) corresponding to latent variables. However, this assumption often fails to hold in many real-world scenarios, thus limiting the identifiability of existing methods. To address this challenge, this study investigates the identifiability of latent variables under the “non-pure” measurement setting and proposes a linear non-Gaussian acyclic latent variable model (LiNGLM) that allows causal relationships among observed variables. Based on this model, a latent variable structure learning (LLSTIN) algorithm is proposed. This algorithm is based on the transformed independent noise (TIN) condition and the graph criterion established thereby. First,

* 基金项目: NSFC-区域创新发展联合基金(广东)(U24A20233)

收稿时间: 2025-10-14; 修改时间: 2025-11-11, 2025-12-26; 采用时间: 2026-01-12; jos 在线出版时间: 2026-05-20

the “single-factor sets” corresponding to each latent variable are extracted through the build causal cluster (BCC) algorithm to identify the existence of latent variables. Then, the “root” observed variables in the sets are selected as effective proxies for latent variables, and the causal relationships between latent variables are further identified. This study theoretically proves that the proposed algorithm can correctly extract single-factor sets and further identify the causal structure among latent variables. Simulation data and real-world data experimental results further verify the correctness and effectiveness of the proposed algorithm.

Key words: causal structure learning; linear non-Gaussian; “impure” measurement scenario; transformed independent noise

因果关系发现是探寻事物间因果联系的研究手段^[1,2]. 在心理学研究中, 从行为观测、问卷数据及实验记录等资料中挖掘变量间真实因果关联, 是揭示心理现象本质、构建理论模型与优化干预策略的关键^[3,4]. 在传统的因果发现任务中, 研究人员专注于发现可测量的变量之间的因果关系, 已经有许多解决此问题的尝试^[5,6]. 而在一些研究中, 科学家们需要推断隐变量之间的因果关系. 例如, 为了研究“开放性”与“外向性”等人格特征之间的关系, 心理学家通过设计相应的调查问卷来间接测量这些感兴趣的潜在因子(即隐变量). 在这种情况下, 如何从隐变量的可观测后代变量(也称测量变量)来推断隐变量间的因果关系成为研究的关键.

在没有任何假设下, 从观测数据推断隐变量的因果结构是一个难以实现的任务. 为此, 研究者通常对模型施加参数化假设. 例如, Silvar 等人^[7]提出了一种线性隐变量模型, 在“3个纯子代”的假设下, 基于 Tetrad 约束, 设计了一种两阶段的隐变量因果结构学习算法: 即通过 Tetrad 约束识别因果聚类, 定位隐变量, 然后基于因果聚类的信息, 间接推断隐变量间的因果结构. Huang 等人^[8]和 Dong 等人^[9]在此基础上, 将该识别性结果推广到线性隐变量层级结构上, 并给出了更一般化的结构学习算法. 尽管这些方法推动了隐变量因果结构的识别性边界, 但是这些方法仅利用观测变量间的二阶统计量信息来推断因果结构, 存在马尔可夫等价类的问题, 即部分隐变量间的因果方向是无法识别的.

为了解决隐变量间的马尔可夫等价类问题, 一种主流的解决方案是引入非高斯假设. Cai 等人^[10]基于非高斯性假设提出 Triad 约束方法, 通过构造测量变量间的伪回归残差并检验其独立性, 首次回答了隐变量间因果方向的识别性问题; Xie 等人^[11]针对 Triad 约束仅适用于单因子模型的局限性进行扩展, 提出广义独立噪声条件 (generalized independent noise, GIN), 使其能够识别变量集之间的因果序. Chen 等人^[12]基于 GIN 条件, 讨论了隐变量因果结构可识别的最小非高斯性条件. 近年来, Xie 等人^[13,14]基于 GIN 条件提出了一种递归学习策略, 回答了线性非高斯隐变量层级结构模型的识别性问题. Jin 等人^[15]讨论了基于 GIN 条件可识别的隐变量因果结构边界. Chen 等人^[16]则进一步探究了基于高阶累积量 (higher-order cumulant) 在存在隐变量影响时对因果结构的识别能力. 然而, 这类研究往往需要满足一定数量的“纯子”假设, 即隐变量对应的观测子代变量(测量变量)间不存在直接的因果关系. 这一假设在许多现实场景中常被违反: 例如在研究大五人格 (big five personality) 中“外向性”与“宜人性”的关系时, 心理学家通过设计对应的问卷问题作为反映“外向性”“宜人性”的观测指标, 而这些问题之间可能存在直接的因果联系. 此时相关算法可能得出错误结论或无法提供有效信息, 给现有方法的可识别性带来了挑战.

针对上述挑战, 本文聚焦于“非纯”测量场景, 提出一种更为一般化的隐变量模型——线性非高斯无环隐变量模型 (linear non-Gaussian acyclic latent variable model, LiNGLM), 该模型放宽了现有的纯子代测量假设^[7-16], 即允许隐变量的测量变量之间存在直接的因果连边(如图1所示). 为了回答 LiNGLM 的结构识别性, 本文引入了变换独立噪声条件 (transformed independent noise, TIN)^[17]及其图准则理论, 基于 TIN 条件, 可以从存在复杂连接关系的观测变量提取各隐变量对应的“单因子集”(见定义5), 并选取单因子集中的“根”观测变量(见第3节)作为各隐变量的有效代理, 进而识别隐变量间的因果关系, 从而提出了 LLSTIN 算法. 本文从理论上证明了 LLSTIN 算法可以在“非纯”场景下正确提取和利用单因子集识别隐变量的因果序, 且仿真数据与真实数据的实验结果进一步验证了该算法的正确性和有效性.

本文第1节介绍相关工作. 第2节阐述本研究所需基础知识, 包括因果模型与问题定义. 第3节介绍因果聚类构建算法. 第4节介绍所提出的基于 TIN 条件的隐变量结构学习算法. 第5节通过仿真实验与真实数据实验验证所提算法的有效性. 第6节针对本文理论结果所依赖的假设展开讨论. 第7节附上相关引理证明. 第8节总结本文及未来拓展方向.

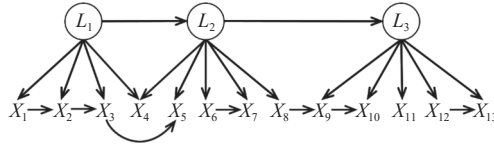


图1 “非纯”场景下本文方法可以处理的示例因果图

1 相关工作

在非时序观测数据的隐变量因果结构学习领域, 主要存在 3 类方法: 基于约束的方法、基于得分的方法以及基于因果函数模型的方法.

基于约束的隐变量因果发现方法以 FCI 算法^[18]为代表, 核心原理是通过扩展独立性和条件独立性检验, 识别观测变量间无法用自身解释的关联模式, 进而推断隐变量的存在及观测变量间因果关系. 然而, 该方法局限在于马尔可夫等价类问题——无法区分满足相同独立性关系的不同因果结构, 仅能输出候选结构集合.

基于得分的隐变量因果发现方法通过设计包含隐变量的评分函数与搜索策略, 寻找最优因果结构. 这类方法的核心是将隐变量纳入模型评分体系. 例如, Silva 等人^[7]提出扩展贝叶斯得分, 通过对隐变量进行边际化处理, 将其视为潜在参数纳入模型评分计算, 以此评估包含隐变量的结构合理性; Ng 等人^[19]则将二阶统计量约束转化为目标优化函数, 设计出基于得分的算法, 通过优化目标函数学习隐变量间的因果结构.

基于因果函数模型的隐变量因果发现方法以线性非高斯无环模型 (linear non-Gaussian acyclic model, LiNGAM)^[20,21]为代表, 其核心原理是假设变量间存在特定函数关系 (如线性、非线性) 并引入独立噪声, 通过分析函数形式与噪声分布 (如非高斯分布) 推断因果方向. 此类方法的优势是可突破部分传统约束方法的局限, 处理更复杂的隐变量关联场景, 但依赖严格的函数形式假设及噪声独立等强条件.

2 因果模型与问题定义

本文旨在“非纯”场景下回答隐变量间的因果结构学习的问题. 首先介绍因果图模型中常用的基本定义和符号系统 (第 2.1 节), 随后正式定义线性非高斯无环隐变量模型 (LiNGLM) 及本文所依赖的假设 (第 2.2 节). 最后, 本文引入变换独立噪声条件及其图准则理论, 该理论结果是本文隐变量因果结构学习算法的关键 (第 2.3 节).

本文涉及的符号及其含义如表 1 所示.

表 1 符号和对应描述

符号	含义
DAG, 简写 G	有向无环图 (directed acyclic graph, DAG)
$\mathbf{V}$	因果图中节点集合
$\mathbf{E}$	因果图中连接各个节点的边的集合
$\mathbf{L}$	因果图中隐变量节点集
$\mathbf{X}$	因果图中观测变量节点集
c	包含 n 个观测变量的集合
$L(c)$	集合 c 代表的隐变量节点
L_i, L_j	$\mathbf{L}$ 中的单个变量 ($i, j = 1, 2, \dots, n$)
X_i, X_j	$\mathbf{X}$ 中的单个变量 ($i, j = 1, 2, \dots, n$)
$V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$	节点 V_i 和节点 V_j 在给定节点集 $\mathbf{S}$ 的条件下相互独立
$\text{Anc}(\mathbf{Z})$	$\mathbf{Z}$ 的祖先节点, 包含 $\mathbf{Z}$ 本身, 表示 $\text{Anc}(\mathbf{Z}) = \{j \mid \exists j \in \mathbf{Z}, j \rightsquigarrow i\} \cup \{\mathbf{Z}\}$

2.1 背景知识

为了从观测数据研究变量间的因果关系, 因果图模型^[22]是一种经典的研究方法. 在因果图模型中, 变量间的

因果关系通常用有向无环图 (DAG) 表示, 该图也称为因果图. 因果图可表示为 $G = \{\mathbf{V}, \mathbf{E}\}$, 其中 $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$ 是节点的集合, $\mathbf{E}$ 为边集. 因果图中, 有向边需满足约束: 若对于 $(V_j, V_i) \in \mathbf{E} (i \neq j)$, 则 $(V_i, V_j) \notin \mathbf{E}$, 此时 V_j 称为 V_i 的父节点, 记为 $V_j \rightarrow V_i$, 边的权重表示 V_j 到 V_i 的因果强度. 一般而言, 在因果图模型中, 需假设因果马尔可夫条件、因果忠诚性假设成立.

假设 1. 因果马尔可夫假设^[22]. 对于因果图 $G = \{\mathbf{V}, \mathbf{E}\}$, 任意 $V_i \in \mathbf{V}$, 给定其所有父节点, V_i 与所有非后代变量条件独立.

假设 2. 因果忠诚性假设^[22]. 对于因果图 $G = \{\mathbf{V}, \mathbf{E}\}$, 任意 $V_i, V_j \in \mathbf{V} (i \neq j)$ 和集合 $\mathbf{S} \subset \mathbf{V}$, 若 $V_i \perp\!\!\!\perp V_j | \mathbf{S}$, 那么 V_i 和 V_j 被集合 $\mathbf{S}$ 中的变量 d-分离.

上述假设保证从数据的统计特性中唯一恢复因果结构的可能性. 其中, d-分离是因果图中判断变量间条件独立性的关键工具, 通过 d-分离可以确定因果结构图中变量之间是否存在直接的因果边. 具体定义如下所示.

定义 1. d-分离 (d-separation)^[22]. 对于因果图 $G = \{\mathbf{V}, \mathbf{E}\}$, 对于 G 上的非邻接节点 $V_i, V_j \in \mathbf{V} (i \neq j)$ 集合 $\mathbf{Z} \subset \mathbf{V}$, V_i 与 V_j 之间的所有连通路径被 $\mathbf{Z}$ 阻断, 则 $\mathbf{Z}$ 在 G 中 d-分离 V_i 与 V_j , 且 $\mathbf{Z}$ 满足:

- (1) 路径上的每个非对撞节点 V_m , 如 $V_i \rightarrow V_m \rightarrow V_j$ ($V_i \leftarrow V_m \leftarrow V_j$), 或者 $V_i \leftarrow V_m \rightarrow V_j$, 都有 $V_m \in \mathbf{Z}$.
- (2) 路径上的每个对撞节点 V_n 与其后代 $\text{Des}(V_n)$, 如 $V_i \rightarrow V_n \leftarrow V_j$, 都有 $V_n \notin \mathbf{E}$, $\text{Des}(V_n) \notin \mathbf{Z}$.

d-分离刻画了两个节点在图上的结构关系, 而 t-分离将其推广到两个节点集之间的结构关系. 在介绍 t-分离之前, 首先回顾一下迹的概念.

定义 2. 迹 (trek)^[23]. 对于因果图 $G = \{\mathbf{V}, \mathbf{E}\}$, 对于 G 上的节点 $V_i, V_j \in \mathbf{V}$, 从 V_i 到 V_j 的一对有序的有向路径 (P_1, P_2) 是一条迹, 其中 P_1 的叶子节点是 V_i , P_2 的叶子节点是 V_j , 且 P_1 和 P_2 共同的源节点为 V_k .

定义 3. t-分离 (trek-separation)^[23]. 对于因果图 $G = \{\mathbf{V}, \mathbf{E}\}$, 设 $\mathbf{M}$ 、 $\mathbf{N}$ 、 $\mathbf{C}_\mathbf{M}$ 、 $\mathbf{C}_\mathbf{N}$ 是 $\mathbf{V}_G$ 的 4 个子集, 它们不必不相交. 若对于从 $\mathbf{M}$ 中的节点到 $\mathbf{N}$ 中的节点的每条路径 (P_1, P_2) , P_1 包含 $\mathbf{C}_\mathbf{M}$ 中的节点, 或 P_2 包含 $\mathbf{C}_\mathbf{N}$ 中的节点, 则称对 $(\mathbf{C}_\mathbf{M}, \mathbf{C}_\mathbf{N})$ 将 $\mathbf{M}$ 与 $\mathbf{N}$ t-分离.

通常来讲, 通过适当的因果假设, 可以从观测数据的分布特性来研究因果图中的“分离”模式 (例如 d-分离、t-分离), 从而进一步确定隐变量的因果结构.

2.2 问题定义

本文的理论结果建立在线性非高斯隐变量模型, 在这个模型中, 因果图中变量之间的因果关系可以形式化为以下公式:

$$V_i = \sum \beta_i Pa_{V_i} + \varepsilon_{V_i} \quad (1)$$

其中, Pa_{V_i} 表示 V_i 的父集, β_i 是从 Pa_{V_i} 到 V_i 的因果效应, ε_{V_i} 表示 V_i 的噪声项, 所有的噪声项都服从非高斯分布, 且不同的噪声项彼此独立. 令 $\mathbf{V} = \mathbf{L} \cup \mathbf{X}$ 表示因果图中的变量集, $\mathbf{L}$ 表示隐变量集, $\mathbf{X}$ 表示观测变量集. 为了不失一般性, 假设 $\mathbf{V}$ 中变量的均值为 0. 本文研究的模型如下所示.

定义 4. 线性非高斯无环隐变量模型 (LiNGLM). 线性非高斯隐变量模型, 其因果图 $G = \{\mathbf{V}_G, \mathbf{E}_G\}$, 定义如下.

- a) $\mathbf{V}_G$ 上的分布满足马尔可夫假设, 并且对有向无环图 G 是忠诚的.
- b) $\mathbf{V}_G = \mathbf{X}_G \cup \mathbf{L}_G$, $\mathbf{X}_G$ 是观测变量的集合, $\mathbf{L}_G$ 是隐变量的集合, $\mathbf{X}_G$ 和 $\mathbf{L}_G$ 中的每个变量都是由公式 (1) 中的结构方程模型生成的.
- c) 在 G 中每对变量之间至少有一条路径.

不同于现有的方法, 如 BPC^[7]、Triad^[10]、GIN^[11]、FOFC^[24]等, 由于这种“非纯”结构的复杂性, 正确识别隐变量的存在极具挑战. 因此, 我们需要寻求“非纯”场景下隐变量可识别的结构条件.

定义 5. 因子集 (factor set). 在 LiNGLM 中, 对于任意 $L_i, L_j \in \mathbf{L}$, 则:

(1) 若存在 $c_i, c_j \in \mathbf{X}$, 其中 $|c_i| \geq 3$, $L(c_i) = L_i$; $|c_j| \geq 3$, $L(c_j) = L_j$, 使得 $c_i \perp\!\!\!\perp c_j | L_i/L_j$, 则 c_i 和 c_j 分别为 L_i 和 L_j 对应的单因子集.

(2) 若存在 $c_j' \in \mathbf{X}$, 使得 $c_j' \not\perp\!\!\!\perp \{c_i, c_j\} | L_i/L_j$, 则 c_j' 为 L_i 和 L_j 包含的多因子集.

对于某一单因子集, 集合中的所有观测变量有且仅有一个隐父亲节点. 在给定该隐父亲节点时, 该集合与其他单因子集分离 (无需与其他所有观测变量分离). 因此, 可以利用单因子集作为隐变量的测量变量, 构建一种“代理回归” (如 GIN、TIN), 通过判断“代理回归”残差与测量变量的独立性不对称性, 来识别隐变量的因果方向. 而多因子集中的观测变量至少存在两个隐变量父代, 即存在某一隐变量到多因子集有不小于两条不相交的迹. 这种结构导致其无法通过单一的隐父亲节点实现与其他单因子集的分离, 这将导致基于 TIN 约束构造的“代理回归”无法“阻塞”来自原因变量的影响, 使得残差与测量变量的独立性不对称性消失了, 无法有效识别隐变量的因果方向. 我们进一步给出如下的例子来进行说明.

以图 1 中 L_2 与 L_3 的因果方向识别为例, 若以 L_2 与 L_3 对应的单因子集变量 X_6 和 X_{11} 作为各自的代理变量, 通过 $TIN(X_6, \{X_6, X_{11}, X_7\}) = 2$, $TIN(X_{11}, \{X_6, X_{11}, X_{12}\}) = 3$ (详见第 4 节), 此时 TIN 条件具有不对称性, 可识别出 L_2 与 L_3 间的因果方向; 若以多因子变量 X_9 作为 L_2 的代理变量, 则通过 $TIN(X_9, \{X_9, X_6, X_{11}\}) = 3$, $TIN(X_{11}, \{X_9, X_{11}, X_{12}\}) = 3$, 此时 TIN 条件不具有不对称性, 无法识别出 L_2 与 L_3 间的因果方向. 基于单因子集的定义, 下面给出本文的稀疏结构约束条件, 这将有助于“非纯”测量场景下隐变量间的方向识别.

条件 1. 稀疏结构约束. 每个隐变量都存在一个单因子集.

条件 1 是保证“非纯”场景下隐变量结构识别性的关键条件之一, 其要求每个隐变量存在含至少 3 个观测变量的单因子集, 且允许观测变量间的直接因果关联. 在现实场景中相较于传统的“纯测量假设”更易满足. 例如, 在设计调查问卷时, 只需为隐变量设计足量观测指标即可, 从而大幅提升了对真实复杂场景的适配性. 关于条件 1 的更多介绍在第 6 节作进一步讨论.

本文所研究的问题为: 在观测变量间存在任意因果关联的“非纯”场景下, 如何设计通用算法分离出单因子集, 并进一步设计算法研究隐变量间的因果关系.

2.3 变换独立噪声条件

在线性非高斯隐变量模型中, 可以利用非高斯性带来的高阶统计量来推断更为复杂的结构信息, 使得“非纯”测量场景下的隐变量识别性有望得到解决. 在本节中, 本文引入一种变换的独立噪声条件, 并介绍该条件建立的统计特性与因果图的关系, 这为后续设计相应的结构学习算法提供了理论保证.

定义 6. 变换独立噪声 (TIN) 条件^[17]. 设 $\mathbf{Z}$ 和 $\mathbf{Y}$ 是观测到的随机变量的两个子集. 假设变量遵循 LiNGAM, 则 TIN 的函数定义如下:

$$TIN(\mathbf{Z}, \mathbf{Y}) := |\mathbf{Y}| - \dim(\Omega_{\mathbf{Z}, \mathbf{Y}}).$$

$TIN(\cdot, \cdot)$ 是一个函数, 其输入为两个随机向量集 $\mathbf{Z}$ 、 $\mathbf{Y}$, 返回值为范围在 $[0, |\mathbf{Y}|]$ 的整数. 其中, $|\mathbf{Y}|$ 表示 $\mathbf{Y}$ 的维数, $\dim(\Omega_{\mathbf{Z}, \mathbf{Y}})$ 表示子空间 $\Omega_{\mathbf{Z}, \mathbf{Y}}$ 的维数, 子空间 $\Omega_{\mathbf{Z}, \mathbf{Y}}$ 是 $\mathbb{R}^{|\mathbf{Y}|}$ 中的一个子空间, 使得 $\omega^T \mathbf{Y} \perp \mathbf{Z}$.

注释: 与独立噪声 (independent noise, IN) 条件^[21]或广义独立噪声 (GIN)^[11]不同, TIN 具有更丰富的表征能力: IN 与 GIN 均可视为 TIN 的特例, 即 $IN(\mathbf{Z}, \mathbf{Y}_i)$ 条件满足当且仅当 $TIN(\mathbf{Z}, \mathbf{Z} \cup \mathbf{Y}_i) = |\mathbf{Z}|$; $GIN(\mathbf{Z}, \mathbf{Y})$ 条件满足当且仅当 $TIN(\mathbf{Z}, \mathbf{Y}) = \text{rank}(\text{cov}(\mathbf{Z}, \mathbf{Y})) < |\mathbf{Y}|$ (其中 $\text{rank}(\text{cov}(\cdot))$ ^[23]表示协方差矩阵的秩).

下面, 我们将给出 TIN 在 LiNGAM 中的图准则, 该准则建立起 TIN 条件与因果图的结构联系.

定理 1. TIN 条件的图表达. 设 $\mathbf{Z}$ 和 $\mathbf{Y}$ 是观测到的随机变量的两个子集, $TIN(\mathbf{Z}, \mathbf{Y})$ 返回的整数值 k 意味着存在一个维数为 k 的集合 $\mathbf{S}$, $0 \leq k \leq |\mathbf{Y}|$. 满足: 1) $\mathbf{S}$ 表示 $\mathbf{Z}$ 的因果效应; 2) $Anc(\mathbf{Z})$ 必须通过 $\mathbf{S}$ 影响 $\mathbf{Y}$; 3) $(\emptyset, \mathbf{S})$ t-分离 $\mathbf{Z}$ 与 $\mathbf{Y}$; 4) $\mathbf{B}_{\mathbf{Y}, Anc(\mathbf{Z})}$ 的秩为 k . 其中, $\mathbf{S}$ 是从 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的最小顶点割集^[17], 确保 G 中没有从 $\mathbf{Z} \setminus \mathbf{S}$ 到 $\mathbf{Y} \setminus \mathbf{S}$ 的有向路径.

例 1: TIN 条件的图表达示例. 在图 1 中, 令 $\mathbf{Z}_1 = \{X_1, X_2\}$ 和 $\mathbf{Y}_1 = \{X_6, X_7\}$, $Anc(\mathbf{Z}_1)$ 包含 $\mathbf{Z}_1$ 所有噪声, 为 $\{X_1, X_2, L_1\}$, 此时, 分离 $Anc(\mathbf{Z}_1)$ 与 $\mathbf{Y}_1$ 的顶点割为 L_2 或 L_1 (从 L_1 到 X_3 及 L_1 到 X_4 的路径均被对撞结构所阻塞), 即 $TIN(\mathbf{Z}_1, \mathbf{Y}_1) = 1$. L_2 包含来自 L_1 的噪声 ε_{L_1} , 该噪声变量表示 $\mathbf{Z}_1$ 的因果效应, 且通过 L_2 传递给 $\mathbf{Y}_1$, 使得 $\{X_6, X_7\}$ 均包含 ε_{L_1} . 在图 1 中即 $(\emptyset, \mathbf{S}_1)$ t-分离 $\mathbf{Z}_1$ 与 $\mathbf{Y}_1$, 其中 $\mathbf{S}_1 \in \{L_1, L_2\}$. $\mathbf{Z}_1$ 中所有的变量到 $\mathbf{Y}_1$ 中所有的变量的路径均需经过 $\mathbf{S}_1$. 此时 $\dim(\Omega_{\mathbf{Z}_1, \mathbf{Y}_1}) = |\mathbf{Y}_1| - TIN(\mathbf{Z}_1, \mathbf{Y}_1) = 2 - 1 = 1$, 即存在 $|\omega| = 1$, 使得 $\omega^T \mathbf{Y}_1 \perp \mathbf{Z}_1$.

基于 TIN 条件的图表达, 可以设计一种测试方案, 通过检验特定观测变量集之间蕴含的 TIN 约束, 来间接推

断隐变量的存在及对应的因果关系. 本文中我们推广 BPC+MIMBuild 的框架 (一种在线性高斯且纯的测量模型场景下的隐变量发现框架)^[7], 通过首先识别因果聚类来推断隐变量的存在性, 而后基于因果聚类为对应的隐变量父亲选择合适的代理变量集来推断隐变量之间的因果关系.

3 因果聚类构建算法

本节针对隐变量对应的观测变量间存在“非纯”关联的问题, 提出一种适用于 LiNGLM 的约束集. 基于该约束集带来的结构信息, 我们提出了一种基于因果聚类构建 (build causal cluster, BCC) 算法, 该方法的核心思路是: 通过 TIN 条件设计一种基于两阶段分离等价类的策略, 使最终得到的等价类能够分离分属不同单因子集的观测变量, 从而正确识别出隐变量的存在.

下面先对聚类算法框架进行概述, 随后在第 3.1–3.4 节中详细阐述算法的实现细节, 并提供必要的理论保证.

如图 2 所示, 因果聚类构建由“子图分离”和“子图搜索”两个阶段构成. 在子图分离阶段, 算法旨在分离不共享同一隐父亲节点集的观测变量, 确保单因子集的分离; 在子图搜索阶段, 则在子图分离得到的子团中进行搜索, 提取各隐变量对应的单因子集. 因果聚类构建将基于观测变量确定隐变量数量, 进而识别隐变量的因果结构. 算法 1 展示了 BCC 的伪代码, 其核心流程如下.

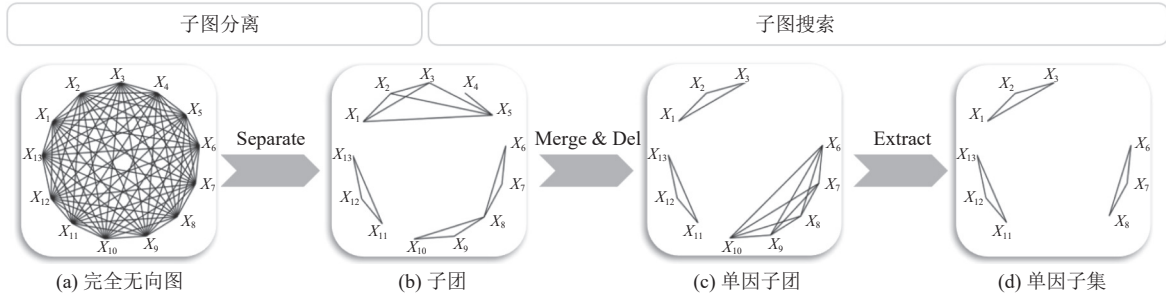


图 2 因果聚类构建框架 (例子为图 1 所示因果图)

在子图分离阶段, 以 Silvar 提出的经典算法 *FindPattern*^[7]为基础, 首先利用 d-分离及基于二阶协方差信息的 Tetrad 约束集 CS1–CS3 对初始图 G 进行初步分离, 得到图 G' (第 1 行), 再通过第 3.1 节提出的引理 1, 确保分属不同单因子集的两个观测变量之间的连边被完全分离, 形成若干子团 (第 2 行); 在子图搜索阶段, 首先执行剪枝操作以辅助单因子集提取: 利用引理 2 合并交点属于单因子集的两个子团 (第 3 行), 再通过引理 3 剔除多因子团, 保证剩余的每个团均包含一个单因子集 (第 4 行), 最终通过单因子集搜索算法提取所有单因子集 (第 5 行).

算法 1. BCC 算法.

输入: 数据 D , 经典算法 *FindPattern*;

输出: 单因子集列表 S_{Sets} .

1. $G' \leftarrow \text{FindPattern}(D)$
2. $\text{Cliques} \leftarrow \text{UseNCS}(G')$ /*通过引理 1 分离 G' 中不同单因子集*/
3. $C' \leftarrow \text{Merge}(\text{Cliques})$ /*通过引理 2 合并交集为单因子集的两个团*/
4. $S_{\text{Cliques}} \leftarrow \text{Del}(C')$ /*通过引理 3 剔除多因子团*/
5. $S_{\text{Sets}} \leftarrow \text{Extract}(S_{\text{Cliques}})$ /*单因子集搜索算法*/

3.1 基于非高斯约束集的子图分离

在子图分离阶段, 算法的核心任务是从因果图中识别并分离分属不同单因子集的变量对, 并在经 *FindPattern*

初步分离得到的 G' 中删除它们之间的连边, 为后续单因子集的提取奠定基础. 以图 2 为例, 从 (a) 到 (c) 展示了真实因果图从初始完全无向图逐步处理至单因子集提取的全过程. 这一阶段的关键挑战在于: 如何以最少的图语言约束为代价, 精准识别并抽离分属不同单因子集的变量对.

为实现这一目标, 本文通过两阶段筛选确定目标等价类, 具体流程如下.

(1) 第 1 阶段: 从图中筛选被隐变量集 d -分离的两个观测变量簇.

(2) 第 2 阶段: 验证两簇的核心变量是否共享隐父亲节点集. 若二者不共享同一隐父亲节点集, 则在完全无向图 G 中删除它们之间的连边, 有助于后续单因子集的提取.

基于上述思路, 我们提出由 3 条 TIN 条件构成的非高斯约束集.

引理 1. 非高斯约束集 (non-Gaussian constraint set, NCS). 对于变量 X_i 和 X_p , 若存在 $X_j, X_q \in \mathbf{X}$ 使得 $TIN(\{X_i, X_j\}, \{X_p, X_q\}) = 1$, 且 $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$, $TIN(\{X_q\}, \{X_i, X_p, X_q\}) = 3$, 则 X_i 与 X_p 不共享同一个隐父亲节点集.

图解释-非高斯约束集能够实现不同单因子集之间的分离. 如图 3 表示不同约束集下的等价类, 其中虚线边代表无向边 ($\leftrightarrow$), 正向边 ($\rightarrow$) 或反向边 ($\leftarrow$); 实线边代表正向边 ($\rightarrow$) 或反向边 ($\leftarrow$); 大括号 ($\{\}$) 表示两个变量等价; 绿色标记对应 TIN 约束下的顶点割. 如图 3(b-4) 所示, 在约束条件 b 下, 存在分属不同单因子集且不共享同一隐父亲节点集的变量等价类. 若单因子集中的变量与多因子集中的变量仅存在一条迹, 引理 1 仍可分离二者, 如图 3(b-1)–(b-3) 所示, 在约束条件 b 下存在两类不共享同一隐父亲节点集的变量等价类. 通过引理 1, 可保证图 G' 中不同单因子集的完全分离, 同时实现部分单因子集与多因子集的分离. 为了确保正文阅读的流畅性, 本文命题以及定理的证明见第 7 节.

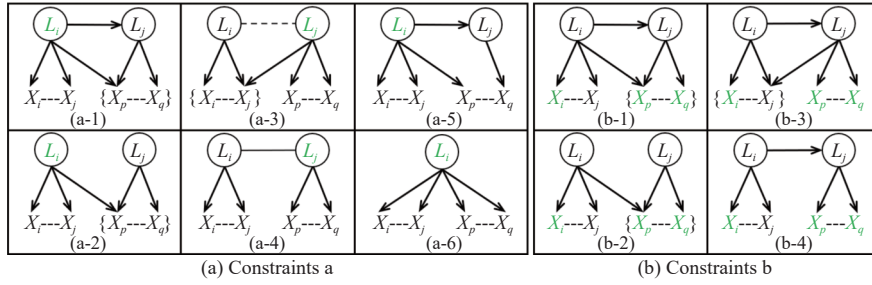


图 3 不同约束集下的等价类

例 2: 非高斯约束集示例. 在图 1 中, 对于分属不同隐父亲节点的变量 X_1 和 X_6 , 二者可被各自的隐父亲节点 L_1 或 L_2 分离, 即存在变量 X_2 和 X_7 , 使得 $TIN(\{X_1, X_2\}, \{X_6, X_7\}) = 1$. 对于 X_6 , 其祖先集 $Anc(X_6) = \{L_1, L_2, X_6\}$, 可得 $TIN(\{X_6\}, \{X_1, X_6, X_7\}) = 3$. 同理 $TIN(\{X_7\}, \{X_1, X_6, X_7\}) = 3$. 综上可判定, X_1 与 X_6 不共享同一个隐父亲节点集.

3.2 基于 TIN 条件的子图搜索

在子图搜索阶段, 算法的核心任务是从团中提取单因子集信息. 为此, 本文提出两条引理 (引理 2、引理 3) 进行剪枝操作, 使最终剩余的每个子团均对应一个单因子集的存在, 且子团的数量对应隐变量的数量 (结论 1), 这有助于单因子集的提取.

多因子集的存在导致团的结构复杂多样, 我们首先对子图分离阶段得到的团进行分析, 并基于团的复杂特征设计团操作规则, 以辅助子图搜索阶段单因子集的提取. 定义如下.

定义 7. 因子团 (factor clique). 对于经过子图分离阶段后得到的子团, 称之为因子团, 其中, 由单因子集组成的团, 称之为单因子团; 由多因子集组成的团, 称之为多因子团; 由单因子集和多因子集共同组成的团, 称之为混合因子团.

由于引理 1 确保不同单因子集之间相互分离, 任意两个单因子集不会构成团, 因此每个单因子团对应唯一的隐变量. 多因子团由两个隐变量下的多因子集构成, 它们彼此间存在至少两条迹而无法通过引理 1 分离. 多因子团

不包含单因子集信息,需予以剔除.对于混合因子团,因其包含的两类变量间存在迹而形成:单因子集的变量存在一条不经过隐变量且指向多因子集变量的迹,导致二者无法通过引理 1 分离.

例 3: 因子团示例. 如图 2(b) 所示, 单因子团 $\{X_{10}, X_{11}, X_{12}\}$ 由 L_3 对应的单因子集所构成; 多因子团 $\{X_4, X_5\}$ 由 L_1 和 L_2 下的多因子集所构成; 对于混合因子团 $\{X_8, X_9, X_{10}\}$, 其中 X_8 为 L_2 下的单因子集变量, $\{X_9, X_{10}\}$ 为 L_2 和 L_3 下的多因子集变量.

根据各类团的性质, 上述 3 类团相交会形成 3 种交集: 第 1 类是单因子团与混合因子团相交, 交集为单因子集; 第 2 类是两个混合因子团相交, 交集为多因子集; 第 3 类是单因子团与多因子团相交, 交集为多因子集. 在子图搜索阶段, 当两团交集为单因子集时, 本文提出引理 2 实现此类团的合并, 以此提取隐变量下的单因子集信息.

引理 2. 合并规则 (merge rule). 对于子团 $C_i, C_j \in C$, 若 $C_i \cap C_j = S_{com}$, 对任意 $X_i \in C_i \setminus S_{com}, X_j \in C_j \setminus S_{com}, X_k \in S_{com}$, 当满足 $TIN(\{X_i\}, \{X_j, X_k\}) + TIN(\{X_j\}, \{X_i, X_k\}) = 3$ 时, 则 S_{com} 为单因子集.

例 4: 合并规则示例. 如图 2(b) 所示, 团 $\{X_6, X_7, X_8\}$ 和团 $\{X_8, X_9, X_{10}\}$ 存在交集 $\{X_8\}$, 通过 $TIN(\{X_6\}, \{X_8, X_9\}) = 1$, $TIN(\{X_9\}, \{X_8, X_6\}) = 2$ 可知 $\{X_8\}$ 属于单因子集. X_6, X_7 与 X_8 同属一个单因子集, 而 X_8 由于与多因子集存在直接的迹而无法通过引理 1 分离, 与 X_9 和 X_{10} 形成的复合团如图 2(c) 所示, 将两个团合并为 $\{X_6, X_7, X_8, X_9, X_{10}\}$.

引理 2 用于判断两个子团的交集是否为单因子集, 若是, 则两团可合并为复合团, 以此聚合同一隐变量对应的单因子集信息, 为后续提取单因子集奠定基础. 若两团的交集为多因子集, 而这类集合无法为隐变量方向的可识别性提供有效信息, 仅需从两个团中分别剔除该交集部分. 同理, 多因子团因不具备有效信息而属于冗余结构, 可通过多因子团剔除规则统一剔除, 如引理 3 所示.

引理 3. 剔除规则 (delete rule). 对于子团 $C_i \in C$, 若对任意 $X_i, X_j \in C_i$, 存在 $C_j \subset C \setminus C_i$, 且对任意 $X_k \in C_j$ 均满足 $TIN(\{X_k\}, \{X_i, X_j\}) = 2$, 则 C_i 为多因子团.

例 5: 剔除规则示例. 如图 2(c) 所示, 团 $\{X_4, X_5\}$ 为隐变量 L_1 和 L_2 下的多因子集. 对于团 $\{X_6, X_7, X_8\}$ 中的任意变量 X_k , 因 $\{L_1, L_2\} \subset Anc(X_k)$, 均满足 $TIN(\{X_k\}, \{X_i, X_j\}) = 2$ (其中 $X_i, X_j \in \{X_4, X_5\}$ (且 $X_i \neq X_j$)), 据此可识别出 $\{X_4, X_5\}$ 为多因子团并剔除.

引理 3 用于识别并剔除多因子团, 从而聚焦单因子集分析. 在基于因果聚类的构建算法中, 经过子图分离与子图搜索阶段对团的合并及剔除操作后, 我们对剩余子图的信息进行归纳总结, 得出推论 1 如下.

推论 1. 子图信息. 从包含所有观测变量的初始完全无向图 G 出发, 通过非高斯约束集 (NCS, 引理 1) 处理后得到图 G' , 包含 $\{C_1, C_2, \dots, C_k\}$ 共 k 个子团, 且满足 $k \geq |\mathbf{L}|$ ($\mathbf{L}$ 为隐变量集合). 进一步通过合并规则 (merge rule, 引理 2) 对符合条件的子团进行合并, 并剔除子团间的多因子交集后, 再经剔除规则 (delete rule, 引理 3) 处理, 最终得到图 G'' , 其中包含 $\{C'_1, C'_2, \dots, C'_m\}$ 共 m 个子团, 此时 $m = |\mathbf{L}|$. 则对任意 $C'_i, C'_j \in G''$, 均满足 $L(C'_i) \cap L(C'_j) = \emptyset$.

在完成团的合并与剔除操作后, 剩余子团包含单因子团与混合因子团, 且二者都包含唯一的单因子集. 基于此性质, 设计单因子集搜索算法如算法 2 所示.

算法 2. 单因子集搜索.

输入: 数据 D , 单因子团 $S_{Cliques}$;

输出: 单因子集列表 S_{Sets} .

1. 初始化单因子集列表 S_{Sets}
 2. **For** $C_i \in S_{Cliques}$, 初始化字典 C_{pair} 为空、计数值 $C_{max} = 0$ 、单因子集 S_{Set} 为空
 3. **For** $(X_i, X_j) \in C_i (i \neq j)$, 初始化计数值 $C_{tin} = 0$
 4. **For** $X_k \in S_{Cliques} \setminus C_i$
 5. **IF** $TIN(\{X_k\}, \{X_i, X_j\}) = 1$
 6. $C_{tin} + 1$
 7. **End if**
-

-
8. $C_{\text{pair}}[(X_i, X_j)] = C_{\text{tin}}$
 9. **End for**
 10. **End for**
 11. $C_{\text{max}} = \max(C_{\text{pair.values}})$
 12. 对于 $C_{\text{pair}}[(X_i, X_j)] = C_{\text{max}}$ 的变量对 (X_i, X_j) , 加入单因子集 S_{Set}
 13. 对 S_{Set} 去重得到单因子集, 并加入单因子集列表 S_{Sets}
 14. **End for**
-

算法 2 展示了单因子集搜索的伪代码, 该算法以单因子团 (见 BCC 算法第 4 行) 作为输入, 对于每个单因子团, 算法遍历团中所有变量对 (X_i, X_j) ($i \neq j$), 若满足 $TIN(\{X_k\}, \{X_i, X_j\}) = 1$, 则计数 C_{tin} 加 1, 并将该计数结果存入变量对计数字典中 (第 1–10 行). 之后, 算法找到变量对计数字典中的最大计数 C_{max} (第 11 行), 并将所有计数等于该最大计数的变量对加入单因子集 S_{Set} (第 12 行). 经过去重处理后, 将得到的单因子集加入单因子集列表 S_{Sets} (第 13 行). 当所有单因子团都处理完毕, 算法输出单因子集列表 S_{Sets} .

例 6: 单因子集搜索示例. 如图 2(c) 所示, 针对团 $\{X_6, X_7, X_8, X_9, X_{10}\}$, 当变量对为 (X_6, X_7) 、 (X_6, X_8) 和 (X_7, X_8) 时, 任取 $X_k \in \{X_1, X_2, X_3, X_{11}, X_{12}, X_{13}\}$, 都满足 $TIN(\{X_k\}, \{X_i, X_j\}) = 1$, 达到最大计数对为 6, 因此将 $\{X_6, X_7, X_8\}$ 纳入单因子集中. 而当变量对包含 X_9 或 X_{10} 时, 仅有团 $\{X_1, X_2, X_3\}$ 满足 TIN 条件, 对应计数对为 3, 故不纳入单因子集. 最终, 识别出的单因子集为 $\{X_1, X_2, X_3\}$ 、 $\{X_6, X_7, X_8\}$ 和 $\{X_{11}, X_{12}, X_{13}\}$, 结果如图 2(d) 所示.

4 基于 TIN 条件的隐变量结构学习算法

通过 BCC 算法从观测变量中提取单因子集后, 即可基于单因子集学习隐变量间的因果序. 首先, 本文通过命题 1 介绍基于 TIN 条件识别隐变量间的因果方向.

命题 1. 基于 TIN 的定向规则. 对于任意 $L_i, L_j \in \mathbf{L}$, 其对应的单因子集为 c_i 和 c_j , 若 c_i 和 c_j 的根观测变量分别为 X_i 和 X_p , 则对于任意 $X_j \in c_i \setminus X_i$, $X_q \in c_j \setminus X_p$, 若 $TIN(\{X_i\}, \{X_i, X_j, X_p\}) = 2$, $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$, 则 $L_i \rightarrow L_j$.

在命题 1 的基础上, 本文设计一种隐变量因果结构学习方法 LLSTIN (learning latent structure based on TIN), 具体如算法 3 所示.

算法 3. LLSTIN 算法.

输入: 数据 D , BCC 算法;

输出: 隐变量因果顺序 K .

1. $S_{\text{Sets}} \leftarrow BCC(D)$ /*(1) BCC 算法输出单因子集*/
 2. **For** $C_i \in S_{\text{Sets}}$, 初始化字典 X_{root} 为空 /*(2) 搜索各单因子集中的根观测变量*/
 3. **For** $X_i \in C_i$
 4. 若对所有 $X_j, X_k \in C_i \setminus X_i$, $TIN(\{X_i\}, \{X_i, X_j, X_k\}) \neq 3$, 则 $X_{\text{root}}[C_i]$ 加入 X_i
 5. **End for**
 6. **End for**
 7. 初始化字典 L_{map} 为空, 将 S_{Sets} 中的子集合映射到隐变量
 8. **WHILE** $\text{len}(L_{\text{map}}) > 0$, 初始化因果序列表 K /*(3) 搜索因果序*/
 9. **For** $L_i \in L_{\text{map}}$
 10. **For** $L_j \in L_{\text{map}} \setminus L_i$
 11. 基于命题 1 定向, 识别 L_i 与 L_j 间的因果方向
 12. **End for**
-

-
13. 若对于 $L_j \in L_{\text{map}} \setminus L_i$, 识别结果均为 $L_i \rightarrow L_j$, 则 $K.add(L_i), L_{\text{map}}.remove(L_i)$
 14. **End for**
 15. **End while**
 16. **Return K**
-

该算法工作流程如下: 首先通过 BCC 算法输出隐变量对应的单因子集 (第 1 行), 其次提取单因子集中的根观测变量, 在一个单因子集里, 根观测变量的祖先集包含了它们共同的隐父亲节点, 不包含其他观测变量, 可作为隐变量的理想代理变量. 基于此性质, 算法设计 TIN 条件识别根观测变量 (第 2–6 行); 并基于每个单因子集初始化隐变量 (第 7 行), 后续基于命题 1 逐步识别隐变量中的外生节点^[20,21], 最终确定隐变量间的因果序 (第 8–15 行).

推论 2. LLSTIN 算法的正确性. 在 LiNGLM 中, 假设条件 1 成立, 则 LLSTIN 算法可以正确识别隐变量及隐变量间的因果序.

例 7: LLSTIN 算法示例. 如图 1 所示, L_1 、 L_2 和 L_3 对应的单因子集分别为 $\{X_1, X_2, X_3\}$ 、 $\{X_6, X_7, X_8\}$ 和 $\{X_{11}, X_{12}, X_{13}\}$, 第 1 步先识别各集合根观测节点, 分别为 $\{X_1\}$ 、 $\{X_6, X_8\}$ 和 $\{X_{11}, X_{12}\}$ (若单因子集全为根观测节点, 此时集合中的变量均为纯测量变量). 接着, 利用根观测变量推导隐变量因果序, 对于 L_1 与 L_2 计算 $TIN(X_1, \{X_1, X_2, X_6\}) = 2$, $TIN(X_6, \{X_1, X_6, X_8\}) = 3$, 识别出 $L_1 \rightarrow L_2$. 同理可识别出 $L_1 \rightarrow L_3$, 由此识别出外生节点 L_1 . 对于 L_2 与 L_3 , 进一步识别出外生节点 L_2 , 最终隐变量的因果序为 $L_1 \rightarrow L_2 \rightarrow L_3$.

4.1 时间复杂度分析

对于 BCC 算法的整体时间复杂度为 $O(n^2m^2)$. 其中 n 为观测变量总数, m 为样本量. 其计算开销集中在子图分离与子图搜索两阶段: 子图分离阶段需遍历所有观测变量对 $O(n^2)$, 每组变量对需验证 3 条 TIN 条件, 每条 TIN 条件依赖 HSIC 检验 $O(m^2)$; 子图搜索阶段的合并规则验证、剔除规则判定及单因子集搜索, 需遍历子图对或变量对 (均 $\leq O(n^2)$), 核心验证步骤仍为 TIN 条件测试, 因此最终整体时间复杂度为 $O(n^2m^2)$.

LLSTIN 算法的整体时间复杂度为 $O(n^2m^2)$. 其中 n 为观测变量总数, m 为样本量. 复杂度由 3 个核心步骤构成: 一是单因子集提取, 直接调用 BCC 算法, 带来 $O(n^2m^2)$ 基础开销; 二是根观测变量识别, 该步骤需遍历所有单因子集 ($l < n$), 每个单因子集内需搜索对应的根观测节点: 对单个观测变量, 需遍历同集内其他变量构成变量对 ($O(n)$), 并通过 TIN 条件测试 ($O(m^2)$), 根观测节点搜索操作时间复杂度为 $O(nm^2)$; 三是因果序推断, 该步骤通过迭代识别外生隐变量确定因果序: 需遍历所有隐变量对 ($l < n$), 每组隐变量对的因果方向判定依赖命题 1 的定向规则, 时间复杂度为 $O(l^2m^2)$. 综合 3 个步骤计算开销, LLSTIN 算法整体时间复杂度为 $O(n^2m^2)$.

综上, BCC 与 LLSTIN 算法通过突破“纯子”假设适配真实复杂场景, 以 $O(n^2m^2)$ 的可控时间复杂度平衡准确性与效率, 既能在中小规模观测变量场景 (如心理学量表、医学诊断指标) 中高效运行, 且在大五人格研究中验证有效性 (见实验部分), 其核心逻辑还可迁移至医学病理因子挖掘、社会科学政策分析等领域, 为依赖隐变量分析的研究与应用提供可靠工具, 具备显著的现实实用性与跨领域影响.

5 实验分析

本节将对本文提出的算法进行实验评估. 第 5.1 节详细介绍实验的对比方法、所用数据集以及评估指标. 第 5.2 和 5.3 节针对算法与其他 4 种代表性的隐变量因果结构学习算法分别在仿真数据和真实数据的实验情况进行分析和评估.

5.1 实验设置

5.1.1 对比方法

在对比方法的选取上, 本文纳入了 3 类具有代表性的线性隐变量因果结构学习算法: 第 1 类是基于线性高斯模型的 ILH 算法^[14], 其核心通过观测变量的秩约束, 实现隐变量的高效定位、基数确定及潜在层次结构识别; 第

2 类是基于线性非高斯模型的方法, 以 GIN^[11] 及其推广版本 PO-LiNGAM^[15] 为代表——PO-LiNGAM 以 GIN 为基础工具, 适用于测量变量与隐变量共同构成部分观测的因果充分线性系统, 且隐变量可处于因果结构任意位置的场景; 第 3 类是基于线性模型研究隐变量层次因果结构的 LHM 算法^[25], 其利用部分非高斯噪声分布数据, 进一步确定隐变量马尔可夫等价类内的部分方向。

上述算法的可识别性均依赖“纯子”假设, 即要求每个隐变量对应至少存在 2-3 个纯测量变量。与之不同的是, 本文聚焦于该理想化假设不成立的场景, 探索隐变量间因果关系的识别方法。

5.1.2 数据介绍

仿真数据: 我们设计了 6 种不同类型的因果场景: Case1 为纯测量场景, 每个隐变量均对应一定数量的纯测量变量; Case2 和 Case3 放松“纯”测量假设, 允许观测变量间存在因果关联; Case4 为完全非纯场景, 引入潜在混淆变量的影响; Case5 和 Case6 则聚焦多因子变量的存在, 探究其在“非纯”场景下对算法可识别性的影响。其中, Case1-Case6 的结构分别如图 4(a)-(f) 所示。

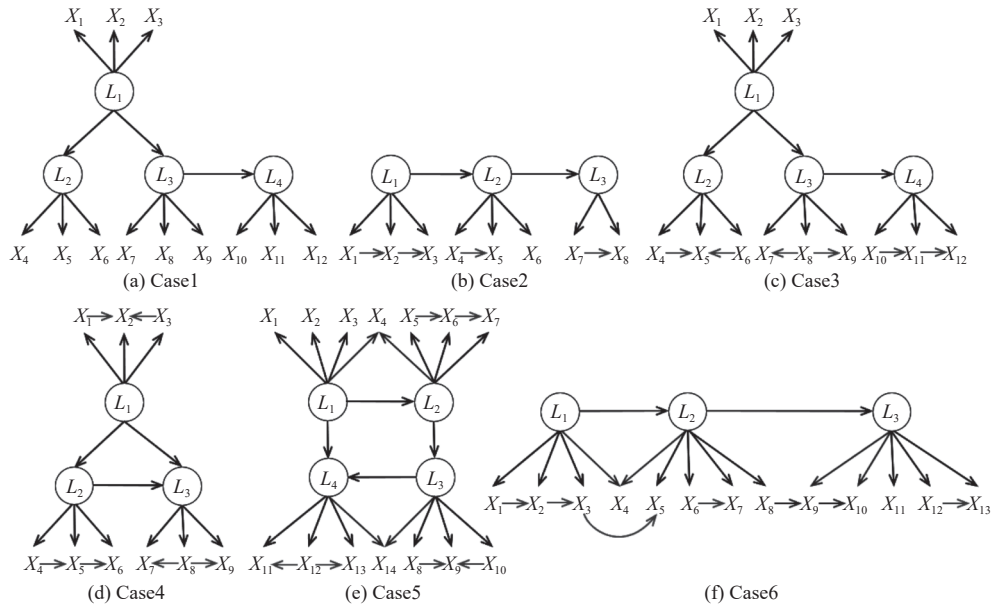


图 4 仿真数据因果图

实验数据基于线性非高斯无环隐变量模型生成 (即公式 (1)), 其中噪声项 ε_{V_i} 服从指数分布, 因果边强度 $\beta_i \sim U(-2, -0.5) \cup U(0.5, 2)$ 表示从父变量 V_j 指向子变量 V_i 的强度。由于数据具有非高斯特性, 故采用希尔伯特-施密特独立准则 (HSIC) 检验^[26] 作为独立测试工具, 显著性水平设为 $\alpha = 0.05$ 。实验中设置了 5k、10k、20k 这 3 种样本量, 通过不同随机种子重复实验 10 次, 最终结果取平均值。

真实数据: 为进一步验证所提方法在心理学实证研究场景中的有效性, 本研究采用基于心理调查问卷采集的大五人格数据集开展实验。大五人格模型, 又称为五因素模型, 由美国心理学家 Costa 和 McCrae^[27] 提出, 是人格心理学领域描述个体人格差异的经典理论框架, 包含开放性 (Openness)、尽责性 (Conscientiousness)、外向性 (Extraversion)、宜人性 (Agreeableness) 和神经质 (Neuroticism), 简称为 O-C-E-A-N。每个维度均对应明确的心理学定义与测量指标, 是心理学研究中人格评估的“黄金标准”之一。

本实验所用数据的采集工具基于国际人格项目库 (IPIP) 的标准化量表开发, 该量表是心理学领域广泛用于大五人格测量的公开工具, 每个维度通过 10 个独立的心理学题项 (调查问卷问题) 实现构念的操作化, 确保测量的信度与效度。具体维度的测量逻辑与问卷题项设计方向如下。

a) 开放性 (Openness): 聚焦个体对新体验、抽象思维的接纳程度, 问卷题项围绕“对抽象概念的兴趣度”“想象力丰富度”“对艺术与创意的敏感度”等核心特征设计, 例如“我常对抽象理论产生兴趣”“我喜欢尝试全新的做事方式”等。

b) 尽责性 (Conscientiousness): 衡量个体的责任感、条理性与目标导向行为, 题项聚焦“任务完成的严谨性”“计划的执行力”“对规则的遵守程度”, 如“我总是提前准备好需要的东西”“我会认真完成每一项任务”。

c) 外向性 (Extraversion): 关注个体的社交倾向与能量来源, 题项围绕“社交活跃度”“热情程度”“寻求刺激的意愿”设计, 例如“我喜欢成为聚会的中心”“我倾向于主动与陌生人交流”。

d) 宜人性 (Agreeableness): 反映个体的人际合作与共情能力, 题项聚焦“对他人的信任度”“合作意愿”“同理心表现”, 如“我愿意为他人提供帮助”“我能理解他人的感受”。

e) 神经质 (Neuroticism): 刻画个体的情绪稳定性, 题项围绕“焦虑倾向”“情绪波动频率”“压力应对方式”设计, 例如“我容易感到紧张不安”“小事也会让我情绪低落”。

数据通过开放心理测量网站 (<https://openpsychometrics.org/>) 的在线互动人格测试平台采集, 该平台是心理学研究领域常用的公开数据收集渠道, 其问卷发放流程符合心理学实证研究的伦理规范 (如知情同意、匿名参与)。本次实验最终获取的数据集包含近 20000 条有效数据, 覆盖 5 个维度共 50 个心理学测量指标 (每个维度 10 个题项)。我们对数据执行标准化处理, 使其均值为 0 且方差为 1。

5.1.3 评估方法

在实验中, 本文将 BCC 算法输出的测量模型与 LLSTIN 算法重建的结构模型^[28], 分别与对比方法进行对比以评估算法的性能。其中, 为评估测量模型的质量, 本文采用以下指标: 潜在遗漏 (Latent omission)、潜在误报 (Latent commission)^[10,11]。其中, OL 为遗漏的隐变量数量, FL 为错误识别的隐变量数量, TL 为真实因果图中的隐变量总数, 计算公式为:

$$Latent\ omission = \frac{OL}{TL}, Latent\ commission = \frac{FL}{TL}.$$

为评估结构模型的重建质量, 本文进一步采用 $F1$ 得分作为评估指标, 其由精确率 (Precision, PRE) 和召回率 (Recall, REC) 组成。精确率反映两个隐变量间重建因果关系中正确估计的比例; 召回率衡量两个隐变量间正确预测的边占网络中实际边的比例。 $F1$ 得分综合考虑精确率和召回率, 可全面评价算法性能。计算公式为:

$$F1 = \frac{2 \times PRE \times REC}{PRE + REC}.$$

5.2 仿真数据实验结果

本节针对仿真数据场景, 对本文所提算法与其他对比方法的实验结果进行分析。其中, BCC 算法输出的测量模型评估结果如表 2、表 3 所示; LLSTIN 算法输出的结构模型评估结果如表 4 所示。其中, 潜在遗漏和潜在误报值越低越好, $F1$ 得分越高越好。具体实验结果分析如下。

在隐变量结构学习任务中, 对比方法中表现最优的为 ILH 算法, 但该算法的性能受场景复杂度显著影响: 随着观测变量间因果关联增强、纯测量变量占比降低, 其输出准确性明显下降。注意到, 对于 Case2, 现有方法的效果均不佳, 这是由于现有的方法要么需要假设 3 个纯的子代 (如 FOFC^[24]), 要么需要非高斯假设且至少两个纯的子代 (如 Triad、GIN^[10,11])。而我们的方法也无法达到理想的性能, 主要原因在于此时条件 1 “每个隐变量都存在一个单因子集”不成立, 其中 L_3 下的观测变量数为 2, 不存在单因子集, 则算法无法检测到该隐变量的存在, 导致效果不佳。相比之下, 本文所提算法在其他实验场景中均保持最佳性能。虽然在测量模型学习的“潜在遗漏”指标评估中, BCC 算法表现略逊于 ILH 算法, 这主要源于 BCC 算法依赖大量 TIN 条件测试, 而独立性测试的准确性对其性能形成了一定限制。而在“非纯”测量场景下, ILH 将观测变量间的因果关联错误建模为额外隐变量的影响, 这种建模方式虽试图适配复杂关联, 但代价是产生更高的潜在误报——其常将实际受单个隐变量影响的观测变量, 错误建模为由多个隐变量复合构成的层级隐变量系统, 导致对隐变量结构的推断偏离真实情况。此外, LHM、PO-LiNGAM 与 GIN 这 3 种方法整体表现不佳, 算法均受限于“纯测量模型”假设, 当该假设不成立时, 算法无法适配观测变量

间的复杂关联, 难以实现隐变量间因果序的有效识别.

总体而言, LLSTIN 算法在仿真数据的隐变量结构学习任务中表现突出, 在所有实验场景下均保持最佳性能, 有效克服了场景复杂度上升、纯测量变量减少带来的性能波动问题. 相较于现有其他算法, 在“非纯”测量场景下 LLSTIN 更能稳定、准确地实现隐变量结构的识别.

表 2 潜在遗漏评估

Case	样本量 (k)	Ours (BCC)	ILH	LHM	PO-LiNGAM	GIN
Case1	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
	10	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
	20	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Case2	5	0.40 (2)	0.00 (0)	0.20 (6)	0.33 (5)	0.10 (6)
	10	0.37 (1)	0.00 (0)	0.20 (4)	0.26 (4)	0.13 (5)
	20	0.33 (0)	0.00 (0)	0.23 (3)	0.13 (2)	0.13 (3)
Case3	5	0.10 (3)	0.00 (0)	—	—	—
	10	0.05 (1)	0.00 (0)	—	—	—
	20	0.05 (1)	0.00 (0)	—	—	—
Case4	5	0.07 (2)	0.00 (0)	—	—	—
	10	0.03 (1)	0.00 (0)	—	—	—
	20	0.00 (0)	0.00 (0)	—	—	—
Case5	5	0.10 (3)	0.15 (2)	—	0.33 (3)	—
	10	0.08 (3)	0.15 (2)	—	0.30 (3)	—
	20	0.05 (2)	0.08 (1)	—	0.30 (2)	—
Case6	5	0.10 (3)	0.00 (0)	—	0.40 (2)	—
	10	0.07 (2)	0.00 (0)	—	0.37 (3)	—
	20	0.03 (1)	0.00 (0)	—	0.33 (2)	—

注: 括号中的数字表示算法无法正确解决问题的次数. 如果一个方法的结果总是错误的, 则用符号“—”表示

表 3 潜在误报评估

Case	样本量 (k)	Ours (BCC)	ILH	LHM	PO-LiNGAM	GIN
Case1	5	0.00 (0)	0.00 (0)	0.13 (0)	0.03 (0)	0.00 (0)
	10	0.00 (0)	0.03 (1)	0.125 (0)	0.03 (0)	0.00 (0)
	20	0.00 (0)	0.00 (0)	0.15 (0)	0.03 (0)	0.00 (0)
Case2	5	0.00 (0)	0.53 (5)	0.47 (6)	0.20 (5)	0.37 (6)
	10	0.00 (0)	0.50 (4)	0.40 (4)	0.22 (4)	0.37 (5)
	20	0.00 (0)	0.43 (2)	0.30 (3)	0.20 (2)	0.40 (3)
Case3	5	0.00 (0)	0.40 (8)	—	—	—
	10	0.00 (0)	0.50 (10)	—	—	—
	20	0.00 (0)	0.45 (8)	—	—	—
Case4	5	0.00 (0)	0.67 (10)	—	—	—
	10	0.00 (0)	0.57 (7)	—	—	—
	20	0.00 (0)	0.53 (6)	—	—	—
Case5	5	0.00 (0)	0.60 (4)	—	0.40 (5)	—
	10	0.00 (0)	0.58 (3)	—	0.38 (6)	—
	20	0.00 (0)	0.60 (4)	—	0.40 (6)	—
Case6	5	0.07 (2)	0.97 (10)	—	0.47 (4)	—
	10	0.03 (1)	0.93 (10)	—	0.50 (5)	—
	20	0.03 (1)	0.93 (10)	—	0.43 (3)	—

注: 括号中的数字表示算法无法正确解决问题的次数. 如果一个方法的结果总是错误的, 则用符号“—”表示

表 4 F1 得分

Case	样本量 (k)	Ours (LLSTIN)	ILH	LHM	PO-LiNGAM	GIN
Case1	5	1.00 (0)	0.93 (0)	0.65 (3)	0.95 (0)	0.94 (1)
	10	1.00 (0)	0.97 (0)	0.70 (2)	0.95 (0)	0.97 (0)
	20	1.00 (0)	1.00 (0)	0.74 (2)	0.98 (0)	1.00 (0)
Case2	5	0.53 (2)	0.30 (5)	0.20 (6)	0.10 (5)	0.25 (6)
	10	0.60 (1)	0.35 (4)	0.33 (4)	0.15 (4)	0.32 (5)
	20	0.67 (0)	0.45 (2)	0.40 (3)	0.25 (2)	0.37 (3)
Case3	5	0.83 (1)	0.70 (4)	—	—	—
	10	0.85 (1)	0.72 (2)	—	—	—
	20	0.90 (1)	0.74 (2)	—	—	—
Case4	5	0.90 (0)	0.56 (6)	—	—	—
	10	0.95 (0)	0.60 (5)	—	—	—
	20	1.00 (0)	0.64 (4)	—	—	—
Case5	5	0.84 (0)	0.30 (7)	—	0.11 (10)	—
	10	0.86 (0)	0.32 (6)	—	0.12 (9)	—
	20	0.92 (0)	0.33 (6)	—	0.15 (9)	—
Case6	5	0.83 (1)	0.30 (8)	—	0.40 (4)	—
	10	0.87 (1)	0.33 (8)	—	0.53 (3)	—
	20	0.92 (0)	0.37 (8)	—	0.60 (2)	—

注: 括号中的数字表示算法无法正确解决问题的次数. 如果一个方法的结果总是错误的, 则用符号“—”表示

5.3 真实数据实验结果

在真实数据集实验中, 我们首先利用 BCC 算法提取表征不同隐变量的单因子集, 结果如表 5 所示. 算法共输出 4 个单因子集, 对应大五人格数据集的 4 大潜在特征, 即开放性、尽责性、外向性、宜人性. 其中, 神经质未提取出单因子集, 推测原因为其潜在特征下直接关联的测量变量仅 1-2 个, 无法构成单因子集.

表 5 大五人格数据集的隐变量及单因子集

Latent	Single factor set
[L ₁] Conscientiousness	[C ₁] I am always prepared. [C ₂] I leave my belongings around. [C ₃] I pay attention to details. [C ₄] I make a mess of things. [C ₅] I get chores done right away. [C ₆] I often forget to put things back in their proper place. [C ₇] I like order. [C ₈] I shirk my duties. [C ₉] I follow a schedule. [C ₁₀] I am exacting in my work.
[L ₂] Agreeableness	[A ₁] I feel little concern for others. [A ₂] I am interested in people. [A ₃] I sympathize with others' feelings. [A ₄] I am not interested in other people's problems. [A ₅] I have a soft heart. [A ₆] I am not really interested in others. [A ₇] I take time out for others. [A ₈] I feel others' emotions.
[L ₃] Openness	[O ₁] I have a rich vocabulary. [O ₂] I have difficulty understanding abstract ideas. [O ₃] I am quick to understand things.
[L ₄] Extraversion	[E ₁] I am the life of the party. [E ₂] I don't talk a lot. [E ₃] I feel comfortable around people. [E ₄] I keep in the background. [E ₅] I start conversations. [E ₆] I have little to say. [E ₇] I talk to a lot of different people at parties. [E ₈] I don't like to draw attention to myself. [E ₉] I don't mind being the center of attention. [E ₁₀] I am quiet around strangers.

进一步分析人格维度与各单因子集的对应关系: 尽责性、宜人性、外向性可清晰界定为大部分对应题项的潜在原因; 而开放性对应的单因子集仅包含 3 个题项 (即 O₁-O₃), 该维度下其他观测变量间可能存在复杂关联, 无法被统一建模为单因子集. 例如, “I have a vivid imagination.” “I am quick to understand things.” “I have a rich vocabulary.” 分别对应想象力、理解力、词汇量. 从心理机制看, 对经验的开放性可通过积极想象与深度反思催生优质想法; 词汇量丰富的个体, 更易开展批判性思考与大量阅读, 这类行为也构成想法生成的重要路径.

在数据中存在一类观测变量: 这类变量仅存在观测父亲节点, 即其因果依赖关系仅源于其他可观测变量, 而无隐父亲节点. 由于有隐父亲节点的观测变量能更直接地表征其对应的隐变量特征, 因此在单因子集提取过程中, 我

们仅聚焦于受隐父亲节点影响的观测变量. 如“*I use difficult words.*”的因果依赖关系源于另一观测变量“*I have a rich vocabulary.*”, 属于“仅含观测父亲节点”的类型; 而“*I have a rich vocabulary.*”不仅对“*Openness*”这一潜在特征的特征更具代表性, 且其不受其他潜在特征影响, 因此将其纳入“*Openness*”潜在特征的单因子集.

得到各隐变量的单因子集后, 即可利用 LLSTIN 算法发现隐变量间的因果结构, 如图 5 所示. 由此可知, 大五人格的各特质并非孤立存在, 尽责性会通过两种方式影响其他特质: 一是因其守规则、重秩序的习惯, 在社交中更易尊重他人边界、配合团队, 进而通过他人反馈显露出更高的宜人性, 比如让人觉得“好相处”; 二是能推动人深入钻研、持续学习, 知识积累多了会拓宽思维, 从而提升对新事物的接受度 (开放性). 同时, 宜人性和开放性也会促进外向性, 前者因看重人际和谐而主动参与群体活动, 多社交会让人慢慢变得更外向; 后者因喜欢探索新奇事物, 愿意突破舒适区去社交, 这种探索欲会直接增强社交活力. 这些特质关联有明确研究支撑, Costa 等人^[27,29]在大五人格模型研究中指出特质间存在直接或间接影响, Tett 等人^[30]的人格特质交互模型也验证, 个体行为会通过社交反馈、主动探索等机制强化特质的外在表现.

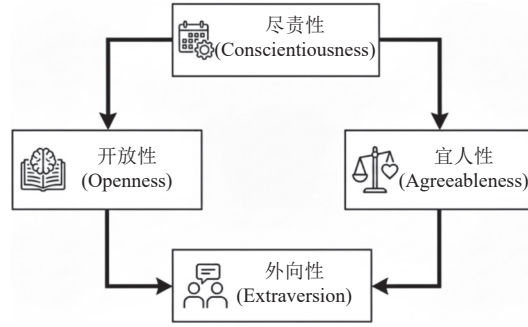


图 5 大五人格数据集算法输出隐变量因果结构

6 讨论

本文突破了传统隐变量因果发现方法依赖“纯测量假设”的局限, 在假设条件 1 成立的情况下, 回答了线性非高斯隐变量模型的识别性. 为了能够推断“非纯”场景下的因果结构, 本文定义了单因子集的概念, 并且要求 $C_i \geq 3$, 这是由于在非纯场景下, 观测变量之间的因果连接影响了这些观测变量作为隐变量的理想测量变量 (例如, 违背了给定隐变量作为条件集时, 观测变量之间条件独立的稀疏性质, 该性质是构造噪声独立约束的关键), 因此, 需要考虑第 3 个变量作为条件变量, 来构造新的条件独立性. 注意到, 在纯的测量假设场景, 单因子集 $C_i = 2$ 就足以推断隐变量的识别性了 (如 Triad、GIN^[11,24]等). 当隐变量对应的观测变量仅 2 个, 且观测变量之间存在直接的因果连接边, 此时, 无法从理论上识别该观测变量集背后的隐变量, 也即, 本文提出的算法无法学到正确的隐变量因果结构.

尽管本模型对单因子集有维数大于 3 的要求, 但在实际应用中, 研究者只需为隐变量多收集一两个观测指标即可满足模型要求; 相比之下, 在复杂现实场景中去强行设计并验证一组完全不存在直接因果关联的“纯测量变量”, 其难度和成本往往远高于增加观测变量的数量. 因此, 本方法在“非纯”测量场景下的应用潜力更为突出.

在可检验性层面, 本文所利用的 TIN 条件可以有效检验存在隐变量影响下观测变量之间的 d-分离关系, 从而在条件 1 成立的前提下, 识别隐变量的存在性, 并推断隐变量之间的因果结构. 当现实场景中, 条件 1 被违背时, TIN 条件仍然能够检验观测变量之间的 d-分离关系, 但是无法从理论上保证隐变量的存在性以及正确推断隐变量之间的因果关系. 因此, 如何放宽条件 1 到更加一般化的场景下, 是本文未来的研究方向之一.

7 理论证明

7.1 引理 1 证明

在证明引理 1 之前, 我们首先介绍: 在线性非高斯隐变量模型中, TIN 条件能够建立起观测变量集 $\mathbf{V}_G$ 中任意

两个子集之间迹的联系, 具体内容如定理 2 所示.

定理 2. LiNGLM 中 TIN 条件的图表达. 设 $\mathbf{Z}$ 和 $\mathbf{Y}$ 是 $\mathbf{V}_G$ 的两个子集, $TIN(\mathbf{Z}, \mathbf{Y}) = k$, 令 $\mathbf{V}_{com} = \mathbf{Z} \cap \mathbf{Y}$, $\mathbf{A} = Anc(\mathbf{Z}) \setminus \mathbf{V}_{com}$, $\mathbf{B} = \mathbf{Y} \setminus \mathbf{V}_{com}$ 则:

(1) 若 $|\mathbf{V}_{com}| = k$, $k < |\mathbf{Y}|$, 则等价于 $\mathbf{A} \perp \mathbf{B} \mid \mathbf{V}_{com}$.

(2) 若 $|\mathbf{V}_{com}| < k$, 则存在 $\mathbf{S}_Z, \mathbf{S}_Y \subset \mathbf{V}_G$, 满足 $|\mathbf{S}_Z| + |\mathbf{S}_Y| = k - |\mathbf{V}_{com}|$, 且 $(\mathbf{S}_Z, \mathbf{S}_Y)$ t-分离 $\mathbf{A}$ 与 $\mathbf{B}$.

证明: (1) 充分性: 由 $\mathbf{V}_{com} = \mathbf{Z} \cap \mathbf{Y}$ 可知 $\mathbf{V}_{com} \subseteq \mathbf{Z} \subseteq Anc(\mathbf{Z})$ 且 $\mathbf{V}_{com} \subseteq \mathbf{Y}$. 因 $TIN(\mathbf{Z}, \mathbf{Y}) = k = |\mathbf{V}_{com}|$, 所有从 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的路径必经过 $\mathbf{V}_{com}$, 即 $\mathbf{V}_{com}$ 是从 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的最小顶点割. 同时, 该最小顶点割阻断了 $\mathbf{A}$ 到 $\mathbf{B}$ 的所有路径, 根据条件独立的图论准则, 可推出 $\mathbf{A} \perp \mathbf{B} \mid \mathbf{V}_{com}$.

必要性: 由 $\mathbf{A} \perp \mathbf{B} \mid \mathbf{V}_{com}$ 可知 $\mathbf{V}_{com}$ 阻断了所有连接 $\mathbf{A}$ 到 $\mathbf{B}$ 的路径, 即 $\mathbf{V}_{com}$ 是从 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的顶点割. 设该顶点割大小为 $|\mathbf{V}_{com}| = t$, 若存在更小的顶点割 $\mathbf{S}' \subset \mathbf{V}_{com}$, 则 $\mathbf{S}'$ 也会阻断所有连接 $\mathbf{A}$ 到 $\mathbf{B}$ 的路径, 与“ $\mathbf{V}_{com}$ 阻断了所有连接 $\mathbf{A}$ 到 $\mathbf{B}$ 的路径”矛盾, 故最小顶点割的大小为 t , 即 $TIN(\mathbf{Z}, \mathbf{Y}) = t$. 结合 $TIN(\mathbf{Z}, \mathbf{Y}) = k$ 可得 $t = k$, 因此 $|\mathbf{V}_{com}| = k$.

(2) 充分性: 由 $TIN(\mathbf{Z}, \mathbf{Y}) = k$ 可知, 存在 $\mathbf{S} \subset \mathbf{V}_G$ 且 $|\mathbf{S}| = k$, $\mathbf{S}$ 是从 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的最小顶点割. 因 $|\mathbf{V}_{com}| < k$, 可将 $\mathbf{S}$ 分解为 $\mathbf{S} = \mathbf{V}_{com} \cup \mathbf{S}_Z \cup \mathbf{S}_Y$ (其中 $\mathbf{S}_Z \subseteq \mathbf{A}$, $\mathbf{S}_Y \subseteq \mathbf{B}$), 故 $|\mathbf{S}_Z| + |\mathbf{S}_Y| = k - |\mathbf{V}_{com}|$. 根据 t-分离准则, 若 $\mathbf{S}$ 是分离 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的顶点割, 则 $(\mathbf{S} \cap Anc(\mathbf{Z}), \mathbf{S} \cap \mathbf{Y})$ 可 t-分离 $Anc(\mathbf{Z})$ 与 $\mathbf{Y}$, 结合 $\mathbf{A} = Anc(\mathbf{Z}) \setminus \mathbf{V}_{com}$, $\mathbf{B} = \mathbf{Y} \setminus \mathbf{V}_{com}$ 可得 $\mathbf{S} \cap \mathbf{A} = \mathbf{S}_Z$, $\mathbf{S} \cap \mathbf{B} = \mathbf{S}_Y$, 因此 $(\mathbf{S}_Z, \mathbf{S}_Y)$ 能 t-分离 $\mathbf{A}$ 与 $\mathbf{B}$.

必要性: 令 $\mathbf{S} = \mathbf{V}_{com} \cup \mathbf{S}_Z \cup \mathbf{S}_Y$, 则 $|\mathbf{S}| = |\mathbf{V}_{com}| + |\mathbf{S}_Z| + |\mathbf{S}_Y| = k$. 因 $(\mathbf{S}_Z, \mathbf{S}_Y)$ t-分离 $\mathbf{A}$ 与 $\mathbf{B}$, 所有连接 $\mathbf{A}$ 到 $\mathbf{B}$ 的路径必经过 $\mathbf{S}_Z$ 或 $\mathbf{S}_Y$, 又因 $\mathbf{V}_{com}$ 阻断 $Anc(\mathbf{Z}) \cap \mathbf{Y}$ 内的路径, 故 $\mathbf{S}$ 阻断了所有连接 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的路径, 即 $\mathbf{S}$ 是顶点割. 若存在更小的顶点割 $\mathbf{S}' \subset \mathbf{S}$ ($|\mathbf{S}'| < k$), 则 $\mathbf{S}' \setminus \mathbf{V}_{com}$ 需阻断所有连接 $\mathbf{A}$ 到 $\mathbf{B}$ 的路径, 与“ $\mathbf{S}$ 阻断了所有连接 $Anc(\mathbf{Z})$ 到 $\mathbf{Y}$ 的路径”矛盾, 故 $\mathbf{S}$ 是最小顶点割, 即 $TIN(\mathbf{Z}, \mathbf{Y}) = k$. 结合 $|\mathbf{S}_Z| + |\mathbf{S}_Y| = k - |\mathbf{V}_{com}| > 0$, 可得 $|\mathbf{V}_{com}| < k$.

接下来, 我们可以开始证明引理 1.

证明: 首先考虑约束条件 a (Constraints a): $TIN(\{X_i, X_j\}, \{X_p, X_q\}) = 1$, 其中 $\mathbf{Z} = \{X_i, X_j\}$, $\mathbf{Y} = \{X_p, X_q\}$. 根据定理 2 (2), 由于 $\mathbf{Z} \cap \mathbf{Y} = \emptyset$, 存在 $\mathbf{S}_Z, \mathbf{S}_Y \subset \mathbf{V}_G$ 且 $|\mathbf{S}_Z| + |\mathbf{S}_Y| = 1$, 使得 $(\mathbf{S}_Z, \mathbf{S}_Y)$ 能 t-分离 $\mathbf{Z}$ 与 $\mathbf{Y}$. 这意味着从观测变量中可提取出两个大小为 2 的簇, 且二者被某一观测变量分离, 这种情况可通过 d-分离直接在完全无向图 G 中移除对应连边. 因此, 此处仅考虑两个簇被隐变量分离的情况 (如图 3(a-1)–(a-6) 所示), 此时 $Anc(\mathbf{Z}) = \{X_i, X_j, L_i\}$. 分离 $\{X_i, X_j, L_i\}$ 与 $\{X_p, X_q\}$ 的顶点割 $\mathbf{S}$ 为 L_i 或 $\{X_p, X_q\}$ 的隐父亲节点 L_j , 且顶点割的大小为 1.

在约束条件 a 对应的等价类中, 图 3(a-1)–(a-4) 中 X_i 与 X_p 均不共享同一隐父亲节点集, 而图 3(a-5) 与 (a-6) 中 X_i 与 X_p 共享同一隐父亲节点集. 为进一步筛选等价类, 在约束条件 a 的基础上加入两条 TIN 条件进行约束, 得到约束条件 b (Constraints b): $TIN(\{X_i, X_j\}, \{X_p, X_q\}) = 1$, 且 $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$, $TIN(\{X_q\}, \{X_i, X_p, X_q\}) = 3$. 以 $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$ 为例, 令 $\mathbf{Z} = \{X_p\}$, $\mathbf{Y} = \{X_i, X_p, X_q\}$, 由定理 2 (2), 因 $\mathbf{Z} \cap \mathbf{Y} = \{X_p\}$, 存在 $\mathbf{S}_Z, \mathbf{S}_Y \subset \mathbf{V}_G$ 且 $|\mathbf{S}_Z| + |\mathbf{S}_Y| \geq 2$, 使得 $(\mathbf{S}_Z, \mathbf{S}_Y)$ t-分离 $Anc(\mathbf{Z}) \setminus \mathbf{V}_{com}$ 与 $\mathbf{Y} \setminus \mathbf{V}_{com}$, 即 $Anc(X_p) \setminus \{X_p\}$ 到 $\{X_i, X_q\}$ 存在大小为 2 的顶点割集. 对于图 3(a-5) 与 (a-6) 分析可知, 若 $Anc(X_p) \setminus \{X_p\} \cap \{X_q\} = \emptyset$ (虚线边代表无向边或正向边, 即 $X_p \leftarrow X_q$ 或 $X_p \rightarrow X_q$), $Anc(X_p) \setminus \{X_p\} = \{L_i\}$, 此时 L_i 到 $\{X_i, X_q\}$ 的顶点割为 L_i 本身, 故 $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 2$, 不满足约束条件; 当 $Anc(X_p) \setminus \{X_p\} \cap \{X_q\} \neq \emptyset$ (虚线边代表反向边, 即 $X_p \leftarrow X_q$), 则 $\{L_i, X_q\} \subseteq Anc(X_p) \setminus \{X_p\}$, 此时 $\{L_i, X_q\}$ 到 $\{X_i, X_q\}$ 的顶点割为 $\{X_i, X_q\}$ 本身, 此时 $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$. 对于 X_p 与 X_q 之间存在的任何关系, 通过双向验证均可拒绝此类等价类.

综上, 通过约束条件 b, 可拒绝等价类图 3(a-5) 与 (a-6); 对于图 3(a-3) 与 (a-4), 部分等价类因不满足约束条件 b 被拒绝. 最终, 约束条件 b 得到的等价类如图 2(b) 所示, 其中 X_i 与 X_p 均不共享同一隐父亲节点集.

7.2 引理 2 证明

证明: 本文通过反证法证明, 假设 $\mathbf{S}_{com}$ 为多因子集. 当 $TIN(\{X_i\}, \{X_j, X_k\}) + TIN(\{X_j\}, \{X_i, X_k\}) = 3$ 时, 由于 $0 < TIN(\{X_i\}, \{X_j, X_k\}) \leq 2$, $0 < TIN(\{X_j\}, \{X_i, X_k\}) \leq 2$. 此时 $TIN(\{X_i\}, \{X_j, X_k\}) = 1$, $TIN(\{X_j\}, \{X_i, X_k\}) = 2$ 或 $TIN(\{X_j\}, \{X_i, X_k\}) = 1$, $TIN(\{X_i\}, \{X_j, X_k\}) = 2$. 由于 $\mathbf{S}_{com}$ 为多因子集, 若 $C_i \setminus \mathbf{S}_{com}$ 与 $C_j \setminus \mathbf{S}_{com}$ 分别只存在一条到 $\mathbf{S}_{com}$ 直接的迹. 此时 $C_i \setminus \mathbf{S}_{com}$

与 S_{com} 、 $C_j \setminus S_{com}$ 与 S_{com} 均可通过引理 1 分离, 则 $C_i \cap C_j = \emptyset$, 与 S_{com} 为多因子集矛盾, 即 $C_i \setminus S_{com}$ 与 $C_j \setminus S_{com}$ 到 S_{com} 分别存在至少两条直接的迹. 对于任意 $X_i \in C_i \setminus S_{com}$ 、 $X_j \in C_j \setminus S_{com}$ 、 $X_k \in S_{com}$, 则有 $TIN(\{X_i\}, \{X_j, X_k\}) = 2$, $TIN(\{X_j\}, \{X_i, X_k\}) = 2$, 与假设矛盾.

当 S_{com} 为单因子集时, 此时 $C_i \setminus S_{com}$ 或 $C_j \setminus S_{com}$ 其中一个为单因子集, 另一个为多因子集. 当 $C_i \setminus S_{com}$ 为单因子集时, 则 S_{com} 存在到 $C_j \setminus S_{com}$ 直接的迹. 若 $C_i \setminus S_{com}$ 存在一条到 S_{com} 直接的迹, 则此时 $C_i \setminus S_{com}$ 与 $C_j \setminus S_{com}$ 无法分离, 与 $C_i \setminus S_{com} \cap C_j \setminus S_{com} = \emptyset$ 矛盾. 则 $C_i \setminus S_{com}$ 不存在一条到 S_{com} 直接的迹, 即 $TIN(\{X_i\}, \{X_j, X_k\}) = 1$, $TIN(\{X_j\}, \{X_i, X_k\}) = 2$. 当 $C_j \setminus S_{com}$ 为单因子集时, 同理可得 $TIN(\{X_j\}, \{X_i, X_k\}) = 1$, $TIN(\{X_i\}, \{X_j, X_k\}) = 2$. 则当 $TIN(\{X_i\}, \{X_j, X_k\}) + TIN(\{X_j\}, \{X_i, X_k\}) = 3$ 时, S_{com} 为单因子集, 证毕.

7.3 引理 3 证明

证明: 本文通过反证法证明, 假设 C_i 不为多因子团. 则 C_i 为单因子团或者混合因子团, 此时 C_i 中至少包含一个单因子集 c_i . 在其他团中, 对于包含单因子集的团, 总存在单因子集变量 X_k , X_k 到 c_i 之间的顶点割为 1. 而对于其余多因子团, 由引理 1 可知 C_i 与其余多因子团是可分离的, 则 C_i 与多因子团变量之间的顶点割也为 1. 则存在 $X_i, X_j \in c_i$, 对于任意 $C_j \subset C \setminus C_i$, 总存在 $X_k \in C_j$, 使得 $TIN(\{X_k\}, \{X_i, X_j\}) = 1$, 与假设矛盾.

当 C_i 为多因子团时, 假定为 L_1 和 L_2 下的多因子集所构成的多因子团. 则总存在 $C_j \subset C \setminus C_i$, 其中 C_j 包含 L_1 或 L_2 的单因子集. 则总存在 $X_k \in C_j$, 使得 $\{L_1, L_2\} \subset Anc(X_k)$, 此时对任意 $X_i, X_j \in C_i$, $Anc(X_k)$ 到 $\{X_i, X_j\}$ 的顶点割为 2, 则有 $TIN(\{X_k\}, \{X_i, X_j\}) = 2$, 则 C_i 为多因子团, 证毕.

7.4 推论 1 证明

证明: 首先, 在子图分离阶段, 通过非高斯约束集 (引理 1) 处理初始完全无向图 G 得到图 G' . 引理 1 表明, 非高斯约束集可识别不共享同一隐父亲节点集的观测变量对, 并在图 G 中删除其连边, 从而使分属不同单因子集的观测变量被分离至不同的子团中. 因此, 图 G' 中的 k 个子团 $\{C_1, C_2, \dots, C_k\}$ 满足: 不同隐父亲节点集对应的观测变量集分属不同子团, 且 $k \geq |\mathbf{L}|$ ($\mathbf{L}$ 为隐变量集合).

其次, 在子图搜索阶段, 对图 G' 中的子团应用合并规则 (引理 2). 引理 2 用于判断两个子团的交集是否为单因子集, 若满足条件则合并为复合团. 由于单因子集属于同一隐变量对应的观测变量集合, 合并操作可将隐变量的单因子集整合, 确保其信息聚合在一个复合团中; 同时, 剔除子团间的多因子交集, 进一步提取单因子集信息.

然后, 应用剔除规则 (引理 3) 剔除多因子团. 多因子团由多因子集组成, 不含单因子集信息, 剔除后剩余子团均包含单因子集.

经过上述操作后, 最终得到 G'' 中的 m 个子团 $\{C'_1, C'_2, \dots, C'_m\}$, 且 $m = |\mathbf{L}|$. 由于每个子团均由同一隐变量对应的单因子集整合而成, 且分离阶段已确保不同子团对应不同隐父亲节点集, 因此任意两个子团 C'_i 和 C'_j 所对应的隐变量集合 $L(C'_i)$ 和 $L(C'_j)$ 必然无交集, 即 $L(C'_i) \cap L(C'_j) = \emptyset$.

7.5 命题 1 证明

证明: 本文通过反证法证明, 假设 $L_j \rightarrow L_i$ 成立. 此时, 分析条件 $TIN(\{X_i\}, \{X_i, X_j, X_p\})$, 由于 X_i 是 L_i 对应单因子集的根观测变量, 其在观测变量间没有父节点, 其祖先节点包括 L_i 和 L_j , 则 $Anc(X_i) = \{X_i, L_i, L_j\}$. 则 $Anc(X_i)$ 到 $\{X_i, X_j, X_p\}$ 的顶点割为 $\{X_i, X_j, X_p\}$ 本身, 则 $TIN(\{X_i\}, \{X_i, X_j, X_p\}) = 3$, 与假设矛盾.

假设 $L_i \rightarrow L_j$ 成立, 分析条件 $TIN(\{X_i\}, \{X_i, X_j, X_p\})$, 由于 X_i 是 L_i 对应单因子集的根观测变量, 其祖先节点包括 L_i , 则 $Anc(X_i) = \{X_i, L_i\}$, $Anc(X_i)$ 到 $\{X_i, X_j, X_p\}$ 的顶点割为 $\{X_i, L_i\}$, $TIN(\{X_i\}, \{X_i, X_j, X_p\}) = 2$. 分析条件 $TIN(\{X_p\}, \{X_i, X_p, X_q\})$, 由于 X_p 是 L_j 对应单因子集的根观测变量, 其祖先节点包括 L_i 和 L_j , 则 $Anc(X_p) = \{X_i, L_i, L_j\}$, $Anc(X_p)$ 到 $\{X_i, X_p, X_q\}$ 的顶点割为 $\{X_i, X_p, X_q\}$ 本身, $TIN(\{X_p\}, \{X_i, X_p, X_q\}) = 3$. 证毕.

7.6 推论 2 证明

证明: 在 LiNGLM 中, LLSTIN 首先依赖 BCC 算法提取单因子集, 而条件 1 保证了每个隐变量存在单因子集,

最终得到的单因子集数量等于隐变量数量且每个单因子集唯一对应一个隐变量; 其次, 算法从单因子集中识别根观测变量, 此类根观测变量的祖先集仅含自身与对应隐变量, 无其他观测变量祖先, 能作为隐变量的有效代理; 最后, 算法通过迭代识别隐变量中的外生变量确定因果序, 结合命题 1 (TIN 定向规则) 可识别隐变量对之间的因果顺序, 识别出其中的外生变量后进行迭代识别, 最终得到隐变量的正确因果序; 因此, LLSTIN 算法可正确识别“非纯”测量场景下隐变量间的因果序, 理论 2 得证。

8 总 结

本文针对隐变量因果发现中传统方法依赖“纯子”假设的局限性, 聚焦于“非纯”测量场景, 提出了线性非高斯无环隐变量模型, 并设计了基于变换独立噪声条件的因果聚类构建 (BCC) 算法与隐变量结构学习 (LLSTIN) 算法。其中, BCC 算法可精准提取各隐变量对应的“单因子集”, LLSTIN 算法则利用集中的“根”观测变量作为代理识别隐变量间因果序。理论证明与仿真、真实数据实验均验证了算法的正确性和有效性, 相较于传统方法, 在“非纯”场景下的隐变量识别准确率显著提升。但研究仍存在模型限于线性关系、高维度场景计算复杂度较高等局限, 未来的研究可以进一步探索非线性隐变量模型拓展与算法效率优化, 以适应更复杂的数据场景。

References

- [1] Spirtes P, Zhang K. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 2016, 3(1): 3. [doi: 10.1186/s40535-016-0018-x]
- [2] Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books, 2018.
- [3] Briganti G, Decety J, Scutari M, McNally RJ, Linkowski P. Using Bayesian networks to investigate psychological constructs: The case of empathy. *Psychological Reports*, 2024, 127(5): 2334–2346. [doi: 10.1177/00332941221146711]
- [4] Steiner PM, Shadish WR, Sullivan KJ. Frameworks for causal inference in psychological science. In: Cooper H, Coutanche MN, McMullen LM, Panter AT, Rindskopf D, Sher KJ, eds. *APA Handbook of Research Methods in Psychology: Foundations, Planning, Measures, and Psychometrics*. 2nd ed., American Psychological Association, 2023. 23–56. [doi: 10.1037/0000318-002]
- [5] Zhao SD, Liu T. Causality and its applications in social media: A survey. *Ruan Jian Xue Bao/Journal of Software*, 2014, 25(12): 2733–2752 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/4724.htm> [doi: 10.13328/j.cnki.jos.004724]
- [6] Cai RC, Chen W, Zhang K, Hao ZF. A survey on non-temporal series observational data based causal discovery. *Chinese Journal of Computers*, 2017, 40(6): 1470–1490 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.01470]
- [7] Silva R, Scheines R, Glymour C, Spirtes P. Learning the structure of linear latent variable models. *The Journal of Machine Learning Research*, 2006, 7: 191–246.
- [8] Huang BW, Low C, Xie F, Glymour C, Zhang K. Latent hierarchical causal structure discovery with rank constraints. *arXiv:2210.01798*, 2022.
- [9] Dong XS, Huang BW, Ng I, Song XC, Zheng YJ, Jin SY, Legaspi R, Spirtes P, Zhang K. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv:2312.11001*, 2023.
- [10] Cai RC, Xie F, Glymour C, Hao ZF, Zhang K. Triad constraints for learning causal structure of latent variables. In: *Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2019. 12883–12892.
- [11] Xie F, Cai RC, Huang BW, Glymour C, Hao Z, Zhang K. Generalized independent noise condition for estimating latent variable causal graphs. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020. 14891–14902.
- [12] Chen ZM, Xie F, Qiao J, Hao ZF, Zhang K, Cai RC. Identification of linear latent variable model with arbitrary distribution. In: *Proc. of the 36th AAAI Conf. on Artificial Intelligence*. AAAI, 2022. 6350–6357. [doi: 10.1609/aaai.v36i6.20585]
- [13] Xie F, Huang BW, Chen ZM, Cai RC, Glymour C, Geng Z, Zhang K. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 2024, 25: 1–61.
- [14] Xie F, Huang BW, Chen ZM, He YB, Geng Z, Zhang K. Identification of linear non-Gaussian latent hierarchical structure. In: *Proc. of the 39th Int'l Conf. on Machine Learning*. Baltimore, 2022. 24370–24387.
- [15] Jin SY, Xie F, Chen GY, Huang BW, Chen ZM, Dong XS, Zhang K. Structural estimation of partially observed linear non-Gaussian acyclic model: A practical approach with identifiability. In: *Proc. of the 2024 Int'l Conf. on Learning Representations*. Vienna: OpenReview.net, 2024.

- [16] Chen W, Huang ZY, Cai RC, Hao ZF, Zhang K. Identification of causal structure with latent variables based on higher order cumulants. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2024. 20353–20361. [doi: [10.1609/aaai.v38i18.30017](https://doi.org/10.1609/aaai.v38i18.30017)]
- [17] Dai HY, Spirtes P, Zhang K. Independence testing-based approach to causal discovery under measurement error and linear non-Gaussian models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 27524–27536.
- [18] Spirtes P, Glymour CN, Scheines R. Causation, Prediction, and Search. 2nd ed., Cambridge: MIT Press, 2000.
- [19] Ng I, Dong XS, Dai HY, Huang BW, Spirtes P, Zhang K. Score-based causal discovery of latent variable causal models. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: JMLR.org, 2024. 37531–37552.
- [20] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. The Journal of Machine Learning Research, 2006, 7: 2003–2030.
- [21] Shimizu S, Inazumi T, Sogawa Y, Hyvärinen A, Kawahara Y, Washio T, Hoyer PO, Bollen K. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. The Journal of Machine Learning Research, 2011, 12: 1225–1248.
- [22] Neuberger LG. Causality: Models, reasoning, and inference, by Judea Pearl, Cambridge University Press, 2000. Econometric Theory, 2003, 19(4): 675–685. [doi: [10.1017/s0266466603004109](https://doi.org/10.1017/s0266466603004109)]
- [23] Sullivant S, Talaska K, Draisma J. Trek separation for Gaussian graphical models. The Annals of Statistics, 2010, 38(3): 1665–1685. [doi: [10.1214/09-aos760](https://doi.org/10.1214/09-aos760)]
- [24] Kummerfeld E, Ramsey J. Causal clustering for 1-factor measurement models. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1655–1664. [doi: [10.1145/2939672.2939838](https://doi.org/10.1145/2939672.2939838)]
- [25] Chen ZM, Xie F, Qiao J, Hao ZF, Cai RC. Some general identification results for linear latent hierarchical causal structure. In: Proc. of the 32nd Int'l Joint Conf. on Artificial Intelligence. Macao, 2023. 3568–3576. [doi: [10.24963/ijcai.2023/397](https://doi.org/10.24963/ijcai.2023/397)]
- [26] Gretton A, Fukumizu K, Teo CH, Song L, Schölkopf B, Smola AJ. A kernel statistical test of independence. In: Proc. of the 21st Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2007. 585–592.
- [27] Costa PT, McCrae RR. The five-factor model of personality and its relevance to personality disorders. Journal of Personality Disorders, 1992, 6(4): 343–359. [doi: [10.1521/pedi.1992.6.4.343](https://doi.org/10.1521/pedi.1992.6.4.343)]
- [28] Xie F, Zeng Y, Chen ZM, He YB, Geng Z, Zhang K. Causal discovery of 1-factor measurement models in linear latent variable models with arbitrary noise distributions. Neurocomputing, 2023, 526: 48–61. [doi: [10.1016/j.neucom.2023.01.034](https://doi.org/10.1016/j.neucom.2023.01.034)]
- [29] McCrae RR, Costa PT. Personality trait structure as a human universal. The American Psychologist, 1997, 52(5): 509–516. [doi: [10.1037/0003-066X.52.5.509](https://doi.org/10.1037/0003-066X.52.5.509)]
- [30] Tett RP, Burnett DD. A personality trait-based interactionist model of job performance. Journal of Applied Psychology, 2003, 88(3): 500–517. [doi: [10.1037/0021-9010.88.3.500](https://doi.org/10.1037/0021-9010.88.3.500)]

附中文参考文献

- [5] 赵森栋, 刘挺. 因果关系及其在社会媒体上的应用研究综述. 软件学报, 2014, 25(12): 2733–2752. <http://www.jos.org.cn/1000-9825/4724.htm> [doi: [10.13328/j.cnki.jos.004724](https://doi.org/10.13328/j.cnki.jos.004724)]
- [6] 蔡瑞初, 陈薇, 张坤, 郝志峰. 基于非时序观察数据的因果关系发现综述. 计算机学报, 2017, 40(6): 1470–1490. [doi: [10.11897/SP.J.1016.2017.01470](https://doi.org/10.11897/SP.J.1016.2017.01470)]

作者简介

王馥弘, 硕士生, CCF 学生会员, 主要研究领域为因果发现.

陈正鸣, 博士, 主要研究领域为因果发现, 潜在因果模型及其应用.

夏业伟, 博士生, 主要研究领域为因果发现, 隐变量模型, 基于核方法的独立性测试.

乔杰, 博士, 主要研究领域为因果推断, 人工智能.

郝志峰, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为算法设计与分析, 数学建模, 数据挖掘.

蔡瑞初, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为因果推断, 深度学习.