一种基于 CDC 的适用于高维数据的因果推断算法

李洪飞[1],万亚平[1,2],阳小华[1,2],耿家兴[1]
(1. 南华大学 计算机学院,湖南 衡阳 421001;
2. 中核集团高可信计算重点学科实验室,湖南 衡阳 421001)

摘 要:一对观测变量之间的因果关系的推断是科学中的基本问题,基于观测数据分析提出因果关系的方法对于产生假设和加速科学发现具有实用价值。利用传统的因果推断算法从高维数据中学习因果网络结构和提高学习准确率是目前研究的难点。在引入耦合相关系数(copula dependence coefficient,CDC)的基础上,提出了一种适用于高维数据的两步骤因果推断算法。首先该算法利用优于最大信息系数的 CDC 对变量间的关联度进行检测,寻找目标节点的父子节点集;然后使用非线性最小二乘独立回归算法,为图中的目标节点与其父子节点之间标注因果方向;最后迭代所有的节点完成完整的因果网络结构。实验结果表明,该算法提高了高维数据下因果网络结构学习的准确率。同时在大样本数据集中,该算法的时间复杂度优于传统算法,对异常值具有鲁棒性。

关键词:耦合相关系数;最大信息系数;最小二乘回归;因果推断
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2020)01-0038-06
doi:10.3969/j.issn.1673-629X.2020.01.007

A High Dimensional Causal Inference Algorithm Based on CDC

LI Hong-fei[1],WAN Ya-ping[1,2],YANG Xiao-hua[1,2],GENG Jia-xing[1]
(1. School of Computer Science,University of South China,Hengyang 421001,China;
2. CNNC Key Laboratory on High Trusted Computing,Hengyang 421001,China)

Abstract:The inference of the causal relationship between a pair of observation variables is a fundamental problem in science. The method of proposing causal relationships based on analysis of observation data is valuable for hypothesis generation and accelerating scientific discovery. It is difficult to study the causal network structure and improve the learning accuracy from high-dimensional data by using traditional causal inference algorithm. Based on copula dependence coefficient (CDC),we propose a two-step causal inference algorithm for high-dimensional data. Firstly,the CDC,which is superior to the maximum information coefficient,is used to detect the degree of correlation between variables for the set of parent and child nodes of the target node. Then,the nonlinear least squares independent regression algorithm is used to distinguish the directions between the target node and its parent and child nodes. Finally all nodes are iterated to complete the causal network structure. The experiment shows that the proposed algorithm improves the accuracy of causal network structure under high dimensional data. At the same time,in the large sample data set,the time complexity of this algorithm is better than that of traditional algorithm,with robustness to outliers.

Key words:copula dependence coefficient;maximum information coefficient;least squares regression;causal inference

0 引 言

因果发现旨在开发从观测中学习因果网络结构的算法,而因果结构学习是机器学习和统计领域的新课题之一[1-2]。包括预防科学在内的许多实证科学中,需要研究各种现象背后的因果机制。

通常,从观测数据中发现因果结构是困难的,因为可能的模型的超指数集,造成一些模型可能无法与其他模型区分开来。在变量 p 数量相当甚至超过样本 n 数量的情况下是复杂的。尽管该问题本身存在困难,但已经开发了许多算法。目前推断算法一般归为两

收稿日期:2019-03-11 **修回日期**:2019-07-15 **网络出版时间**:2019-09-25
基金项目:中央军委科技委创新特区项目(17-163-15-XJ-002-002-04);国家自然科学基金(11805093);湖南省教育重点项目(17A185);湖南省自然科学基金资助项目(2019JJ0486)
作者简介:李洪飞(1992-),男,硕士研究生,研究方向为机器学习、因果关系发现;万亚平,教授,硕导,CCF 会员(14108M),研究方向为大数据与因果关系;阳小华,教授,博导,研究方向为大数据与因果关系、舆情分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190925.1525.070.html

类:贝叶斯结构学习算法[3-5]和基于加性噪声模型因果发现算法[6-9]。例如 PC 算法[3]是一种基于约束的贝叶斯结构学习算法,它先推断出一组条件独立关系,然后识别相关的马尔可夫等价类。但这类方法无法区分 $X \to Z \to Y$ 和 $X \leftarrow Z \leftarrow Y$ 两种结构。加性噪声模型中,在各种假设下,可从观测数据中恢复准确的图像[10]。如 Shimizu 等假设数据遵循线性非高斯无环模型(LiNGAM)[7],后续的 DirectLiNGAM[8] 和 PLiNGAM[11]方法通过两两统计量迭代选择因果顺序。LiNGAM 提出的这些方法在高维设置中是不一致的,这些设置允许变量数的伸缩与样本大小相同或更快,并且只在线性模型下适用。Hoyer 针对非线性数据因果模型,提出适用于连续数据算法 ANM[6]。为扩展到离散数据的情况,对 ANM 算法进一步改进,提出后非线性数据的 PostNonLinear 算法[9]和信息几何方法[12]。

近年来,Makoto 等提出最小二乘独立回归算法(LSIR)[13]。该算法通过交叉验证自然地优化内核宽度和正则化参数等调优参数,从而避免以依赖数据的方式过度拟合。这些方法都只考虑了少量的变量,一旦在高维的情况下($n>7$),因果发现方法的输出可能高度依赖于顺序,准确率会很低。在大多数情况下是两个变量进行方向识别。因为根据理论和算法,可扩展性有限,所以这些因果推断方法都不适用在高维数据的网络结构中。

为解决在高维数据下学习到较准确的因果网络结构问题,曾千千等提出基于 MIC 和 MI 的因果推断算法[14-15]。这些算法将高维网络结构学习问题分解成网络中每个节点对应的低维因果网络结构学习问题。这些算法中,基于信息论的互信息(MI)或者最大信息系数(MIC)能够较好地删除不相关的节点,降低计算复杂度,但 MIC 对于高维变量不可用,对于大样本问题耗时 MI 也不稳定。随着样本量的增大,方差没有变小,对异常值的鲁棒性较弱。由 Jiang 等提出的 CDC 相比较互信息和最大信息系数,可更为准确地检测出变量之间的关联关系[16]。

基于此,文中提出一种基于 CDC 的高维数据因果推断算法。该算法利用 CDC 对变量间的依赖度进行检测,再利用条件独立性检测精炼目标节点的父子节点集,然后使用非线性最小二乘独立回归(LSIR)(两个变量之间效应方向的新测度),为图中的目标节点及其父子节点之间的无向边标注因果方向,最后迭代所有的节点完成完整的因果网络结构。由于 CDC 对大样本下异常值的鲁棒性强,因此提高了算法的准确率。实验也表明在高维数据下该算法的精确度优于其他因果推断算法。

1　预备知识

1.1　加性噪声模型

观察变量为 x_i, $i=\{1,2,\cdots,n\}$, Cause→Result 即, x_j 可以由 $x_i(i<j)$ 完全决定,且不依赖于 $x_l(l>j)$,此因果序列可用 k_i 表示。用递归的方式可以生成该变量,每个模型都可以表示成一个 DAG。每个观察变量 x_i 可用非线性函数表示, $x_i=\sum_{k(j)<k(i)} f(x_j)+e_i$,其中 f 是非线性函数, e_i 表示噪声变量。噪声项 e_i 是具有连续值的随机变量,服从方差非零的分布,且 e_i 之间相互独立,即 $p(e_1,e_2,\cdots,e_{m_e})=\prod_i p_i(e_i)$。

1.2　最小二乘独立回归

最小二乘独立回归(LSIR)是通过最小化输入和残差之间的平方损失互信息估计值来学习加性噪声模型,识别两个非线性变量的因果关系的方法。LSIR 相对现有方法的一个显著优势是,调优参数(如内核宽度和正则化参数)可以通过交叉验证自然地进行优化,从而避免以依赖数据的方式过度拟合。LSIR 与最先进的因果推理方法相比是比较有利的。

假设随机变量 X 和 Y 有附加噪声模型[6]: $Y=f(X)+E$,其中 $f:R\to R$ 是非线性函数, $E\in R$ 是一个零均值与随机变量 X 独立。从一对样本 $\{(x_i,y_i)\}_{i=1}^n$ 中,利用依赖最小化回归 $f_\beta(x)=\sum_{l=1}^m \beta_l\psi_l(x)=\beta^{\mathrm{T}}\psi(x)$ 得到函数 $\hat{f}$,使输入 X 和估计的加性噪声 $\hat{E}=Y-\hat{f}(X)$ 相互独立。在依赖最小化回归中,学习回归参数 β:

$$\min_\beta[I(X,\hat{E})+\frac{\gamma}{2}\beta^{\mathrm{T}}\beta] \tag{1}$$

其中, $I(X,\hat{E})$ 是 X 和 $\hat{E}$ 之间独立性的度量; $\gamma\geqslant 0$ 是避免过度拟合的正则化参数。

采用平方损失相互信息(SMI)[14]作为独立措施:

$$\mathrm{SMI}(X,\hat{E})=\frac{1}{2}\iint\left(\frac{p(x,\hat{e})}{p(x)p(\hat{e})}-1\right)^2 p(x)p(\hat{e})\mathrm{d}x\mathrm{d}\hat{e} \tag{2}$$

其中, $\mathrm{SMI}(X,\hat{E})$ 是从 $p(x,\hat{e})$ 到 $p(x)p(\hat{e})$ 的皮尔逊散度,当且仅当 $p(x,\hat{e})$ 与 $p(x)p(\hat{e})$ 一致时,即 X 和 $\hat{E}$ 是独立的。

使用 SMI,获得一个分析形式的估计值。得到 SMI 估计量后,开始学习参数 β 回归模型:

$$\hat{\beta}=\mathrm{argmin}_\beta[\widehat{\mathrm{SMI}}(X,\hat{E})+\frac{\gamma}{2}\beta^{\mathrm{T}}\beta] \tag{3}$$

这种方法称为最小二乘独立回归。

对于回归参数学习,可以简单地采用梯度下降法:

$$\beta\leftarrow\beta-\eta\left(\frac{\partial\widehat{\mathrm{SMI}}(X,\hat{E})}{\partial\beta}+\gamma\beta\right) \tag{4}$$

其中，η 是步长，可以选一些近似线搜索方法，如 Armijo 规则[14]。

$\mathrm{SMI}(\mathrm{d}x,\hat{E})$ 的偏导数对 β 可以近似表示为：

$$\frac{\partial \mathrm{SMI}(\hat{X},\hat{E})}{\partial \beta} \approx \sum_{l=1}^{b} \hat{\partial}_l \frac{\partial \hat{h}_l}{\partial \beta} - \frac{1}{2}\sum_{l,l'=1}^{b} \hat{a}_l \hat{a}_{l'} \frac{\partial \hat{H}_{l,l'}}{\partial \beta} \tag{5}$$

上面的推导忽视了 β 在 $\hat{e}_i$ 的依赖。使用这个近似表达式具有计算效率。假设噪声 E 的均值为零。考虑到这一点，则将回归量修改为：

$$\hat{f}(x) = f_\beta(x) + \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\beta(x_i)) \tag{6}$$

LSIR 包含三个调优参数的基函数，内核宽度 τ，正则化参数 γ（令 $m = \min(200, n)$），用网格搜索的 CV 选择 τ 和 γ）。具体地，首先对 $\{(x_i,y_i)\}_{i=1}^{n}$ 运行 LSIR，得到一个回归函数 $\hat{f}$。该过程提供了 $\{(x_i,\hat{e}_i) \mid \hat{e}_i = y_i - \hat{f}(x_i)\}_{i=1}^{n}$ 的 SMI 估计值。随机地将输入对和残差 $\{(x_i,\hat{e}_i)\}_{i=1}^{n}$ 置换为 $\{(x_i,\hat{e}_{k(i)})\}_{i=1}^{n}$，$k(\cdot)$ 是一个随机生成的置换函数。置换后的样本对彼此独立，因为随机置换打破了 X 和 $\hat{E}$ 之间的依赖关系（如果存在的话）。然后，计算 LSMI 对被打乱的数据 A 的 SMI 估计。这种随机排列过程被重复多次（在实验中，重复次数设置为 1 000），在零假设（即独立）是建构的。最后，通过评估从原始输入和残差数据计算出的 SMI 估计值相对于随机排列数据的 SMI 估计值分布的相对排序来近似 p 值。

为了确定因果关系的方向，计算两个方向 $X \to Y$（即 X 引起 Y）和 $X \leftarrow Y$（即 Y 导致 X）的 p 值 $p_{X\to Y}$ 和 $p_{X\leftarrow Y}$。对于一个给定的显著性水平 δ，确定的因果方向如下。

如果 $p_{X\to Y} > \delta$ 和 $p_{X\leftarrow Y} \leqslant \delta$，模型选择 $X \to Y$；

如果 $p_{X\leftarrow Y} > \delta$ 和 $p_{X\leftarrow Y} \leqslant \delta$，$X \leftarrow Y$ 模型被选中；

如果 $p_{X\to Y}, p_{X\leftarrow Y} \leqslant \delta$，$X$ 和 Y 之间没有因果关系；

如果 $p_{X\to Y}, p_{X\leftarrow Y} > \delta$，则建模假设是不正确的。

总之，检验因果模型 $Y = f_Y(X) + E_Y$ 或替代 $X = f_X(Y) + E_X$ 是否与数据吻合较好，其中拟合优度是通过输入与残差之间的独立性来衡量的（即噪声估计），输入和残差的独立性可以通过置换测试在实践中确定[14]。

1.3 Copula 相关系数

Copula 相关系数（CDC）用来测量两个随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的依赖关系，是一种较好的关联检测鲁棒依赖测度。CDC 对异常值更具有鲁棒性，适用于更广泛的模型，尤其适用于高维问题。

定理 1（概率积分变换）：对于具有累积分布函数（CDF）F 的连续随机变量 X，随机变量 $U = F(X)$ 在 $[0,1]$ 上均匀分布。因此，向量：

$$\boldsymbol{U} = (U_1,U_2,\cdots,U_p) = P(X) = (F_1(X_1),F_2(X_2),\cdots,F_p(X_p)) \tag{7}$$

称为 Copula 变换，具有均匀的边缘。

定理 2：让随机向量 $\boldsymbol{X} = (X_1,X_2,\cdots,X_p)$ 连续边际 CDFs，$F_i, 1 \leqslant i \leqslant p$。然后 X 的联合累积分布函数 $F(X), X = (x_1,x_2,\cdots,x_p)$，唯一表示为：

$$F(X) = Cx(F_1(x_1),F_2(x_2),\cdots,F_k(x_p)) \tag{8}$$

其中分布函数 C_x 称为 X 的 Copula。

定理 3：设 X 和 Y 为连续随机变量，那么 X 和 Y 是独立的，当且仅当 $C_{XY}(F_X(x),F_Y(y)) = F_X(x)F_Y(y)$，其中 $F_X(x)$ 和 $F_Y(y)$ 分别为 x 和 y 的分布函数。

首先考虑特殊情况 $p = q = 1$。如果 X 和 Y 是独立的，将 $F_X(X)\text{and}\ F_Y(Y)$ 分别表示为 X 和 Y 的分布函数，由定理 3 可知：

$$C_{XY}(F_X(x),F_Y(y)) = F_X(x)F_Y(y) \tag{9}$$

则测量 X 和 Y 之间的依赖关系，可以测量从估计的 $C_{XY}(F_X(x),F_Y(y))$，$\hat{C}_{XY}(F_X(x),F_Y(y))$ 到 $F_X(x)F_Y(y)$ 的距离，这就产生了 Hoeffding－type 统计[13]。但当 $p,q > 1$ 时，Hoeffding－type 的统计量很难构造，因此需要使用其他方法来测量随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的依赖关系。

如果 X 和 Y 是独立的，有 $W = F_X(X)\ \text{and}\ V = F_Y(Y)$ 是独立的。因此，将多维依赖检验问题转化为零假设（H_0），W 和 V 是独立的二维依赖检验问题，可以通过 MIC、互信息等提出的另一种依赖测度来求解。

定义 1：设 X 和 Y 为两个随机变量，X 和 Y 的最大相关系数（MCC）为：

$$\mathrm{MCC}(X,Y) = \sup_{\varphi_1,\varphi_2}\rho(\varphi_1(X),\varphi_2(Y)) \tag{10}$$

其中，$\rho(X,Y)$ 是 X 和 Y 之间的皮尔逊系数，φ_1，$\varphi_2 \in \zeta_2, \zeta_2 = \{\varphi : \int\varphi^2 < \infty\}$。

如果将 φ_1,φ_2 限制为线性函数，MCC 是皮尔森相关系数。为了计算 MCC，通常会将限制 φ_1,φ_2 限制为再生核希尔伯特空间理论（RKHS），例如 n 阶的多项式函数的空间。而 ACE 用来计算 MCC。

定义 2：利用上述表示法，两个随机向量 X 和 Y 之间的 CDC 由 $F_X(X)\ \text{and}\ F_Y(Y)$ 之间的 MCC 给出：

$$\mathrm{CDC}(X,Y) = \mathrm{MCC}(F_X(X),F_Y(Y)) \tag{11}$$

其中，$F_X(x)\ \text{and}\ F_Y(y)$ 分别为 X 和 Y 的分布函数。

可以看出，CDC 是一个基于秩的依赖性测度，因为分布函数是基于秩的，所以 CDC 适用于高维变量。在实际应用中，真实的边际分布函数是未知的，用经验边际分布或估计边际分布代替，得到 CDC 的估计。

2　因果结构学习算法

本节根据相较于互信息和最大信息系数更为准确地检测出变量间关联关系的 CDC,结合发现因果骨架的结构学习,二元变量的方向识别,最终迭代得到一个完整表现出高维数据间因果关系的因果网络结构。研究表明,张等基于互信息与曾等基于最大信息系数的算法构成的无向图,对大样本和高维下异常值的鲁棒性不高,一定程度会增加父子节点集,造成条件独立测试的计算复杂度,得到的最终网络与真实网络差异很大[14-15]。相较于 MIC 和 MI,CDC 能够更好地度量节点间的依赖关系,减少冗余带来的计算复杂度,从而提高效率,并获得更加准确的因果网络结构。该算法基本框架如图 1 所示。

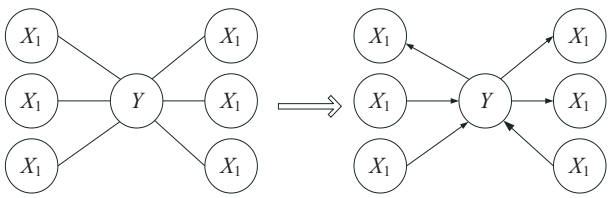


图 1　算法的基本框架

2.1　无向子图构造

CDC 可以快速对随机变量之间的依赖关系进行检测。在样本量增多的情况下,CDC 是最佳的依赖测度,它对异常值具有鲁棒性,计算效率高,对大多数函数类型具有较好的适用性。对随机变量 X 和 Y 计算 CDC(X,Y), X 和 Y 之间的依赖程度与 CDC 的值成正比。当 CDC 值越大,则两个变量对应的点相连程度越高。反之,若 X 和 Y 之间 CDC 值很小,甚至达到 0,则两个变量之间是独立的,对应的两点不相连。根据 CDC 的优点计算变量间的 CDC(X,Y) 值构造无向图,步骤如下:

(1)对数据集中任意的两个随机变量计算其之间的 CDC 值,找到每个节点与其他节点的 CDC 值中最大的值,用 MCDC(X,Y) 表示;

(2)对于任意节点,假设为 y ,其他节点表示为 X 集,如果满足下列条件,则直接连接在变量 X 和 Y 之间建立: $\mathrm{CDC}(x_i,y) \geqslant \alpha \cdot \mathrm{MCDC}(X)$,否则将 x_i 从 X 中移除;

(3)将任意的 $x_i \in X$ 加入到 PC(y),应用条件独立性测试(CI)删除 PC(y)中非父子节点;

(4)重复步骤 3,迭代完 X 中所有节点。

上述步骤中, $0.5 \leqslant \alpha \leqslant 1$ 是一个参数,来确定此阶段的连接数量,如果该参数非常高(接近 1.0),在该算法的这个阶段,存在多个真边被拒绝的高概率。另一方面,将该参数设置的低,可能会导致在算法的这个阶段包含几个错误的边。对于该研究中的问题,参数 α 设置为 0.9,并且发现包括大多数真实边缘,同时允许在网络结构中仅添加少量错误边缘。

2.2　目标节点和父子节点集 PC(y)中每个节点进行方向识别

在前一阶段用 CDC、CI 测试得到的父子节点集,构成了一个准确的骨架。再利用 LSIR 算法对骨架之间每两个节点之间的无向边进行方向识别。相比较于 ANM、IGCI 二元算法,LSIR 在大样本下通过交叉验证进行优化,避免了过分依赖数据而导致的过度拟合。虽然 LSIR 算法不能有效处理高维数据下的因果网络结构学习问题,但由于在算法 2.1 无向图构造过程中,已经得到了目标节点 y 的父子节点集合 PC(y),所以可以利用 LSIR 对 PC(y)中的每一个变量和 y 之间进行方向判别,这等同于在 2 维间应用 LSIR。这种分治策略保证了它的有效性。

2.3　算法描述

文中算法记为 CDC & LSIR causal inference (CLCI)。

Input: variable set $X = \{x_1, x_2, \cdots, x_n\}$; threshold α

Output: DAG G

1 for each $x_i \in X$ do

Set $x_i = y$, $x_i \in X$, PC(y) = { } / * 选取 X 中任意一个节点为目标节点 * /

2 for each $x_i \in X$ do / * 对 PC(y)进行精炼简化,除掉非父节点 * /

if $\mathrm{CDC}(y; x_i) < \alpha * \mathrm{MCDC}(X)$,

then $X = X \setminus x_i$

3 for each $x_i \in X$ do / * 构造关于节点 y 的无向图 * /

add x_i to PC(y), if there is a set S (arbitrary subset of PC (y) $\setminus x_i$),

such that $y \perp x_i \mid S$, then PC(y) = PC(y) $\setminus x_i$

4 for each $x_i \in$ PC(y) do

if there is a set S (arbitrary subset of PC(y) $\setminus x_i$), such that $y \perp x_i \mid S$, then PC(y) = PC(y) $\setminus x_i$

5 for each $x_i \in$ PC(y) do / * 对骨架之间每两个点的方向进行判别 * /

employs LSIR algorithm to distinguish the direction of (y , x_i)

3　模拟数据实验结果与分析

3.1　实验设置与环境

通过人工合成模拟数据进行实验,实验一用于在大样本下测试 CDC 的性能,而实验二用于验证 CLCI 算法的有效性。实验中尽可能假设复杂条件,即观察样本数目足够大,并且模糊了对噪声变量非高斯性的假设。每个节点在因果网络结构中的序列按照函数 $y = \omega_1 * \tan(x_1) + \omega_2 * \tan(x_2) + \varepsilon$ 生成。每个函数中

的权值分别为 ω_1、ω_2，随机取值 0.3 ~ 0.7 之间，y 的父节点分别为 X_1 和 X_2。噪声 ε 服从均匀分布，且权值为 6%。文中提出的算法记为 CLCI。实验平台为 Window 7 64 bit 操作系统，配置为 Intel 酷睿 i5 - 4200U，内存 4 GB，实验环境为 Matlab-R2016a。

3.2　实验分析

在不同样本下，CDC 与 MI 和 MIC 计算时间对比如图 2 所示(表示时间效率)。

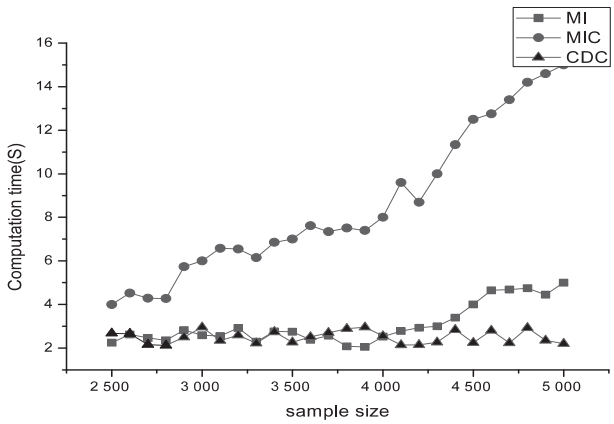
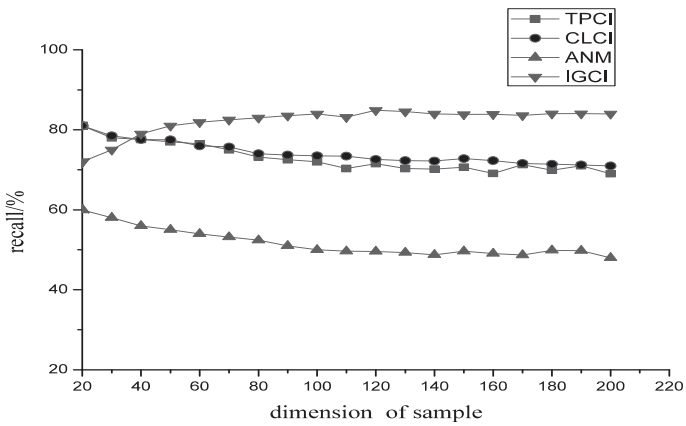


图 2　MI、MIC 和 CDC 在不同大样本下的计算时间对比

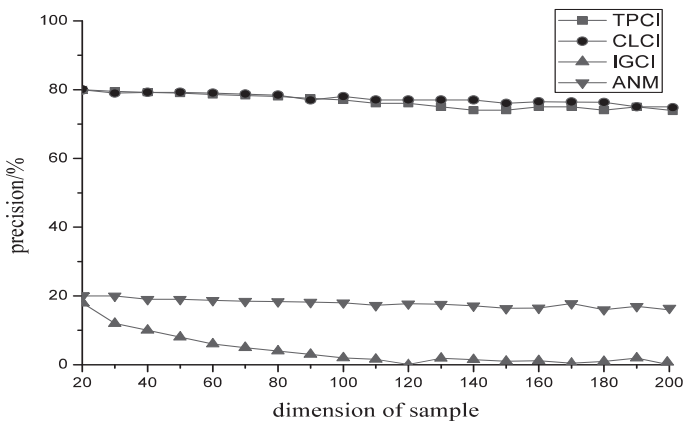
从结果可以看出，随着样本量的增大，MIC 的运行时间迅速增加，MI 直到样本大小约为 4 500 时才开始缓慢增加，而 CDC 的运行时间则没有增加趋势。

为评价 CLCI 在高维下的优劣，避免随机性，在实验中选择 5 000 样本量分别生成 20 维到 200 维的因果网络图，用来测试高维度下不同样本量的实验效果。实验引入三个常用指标 recall、precision 和 F_1 来评价算法性能。在不同维度下四种算法的评分参数如图 3 所示。

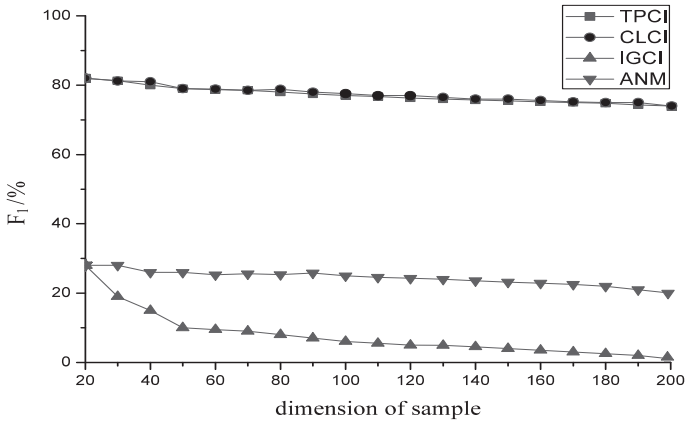
从图 3 可以看到，IGCI 算法的召回率在超过 40 维左右时比其他三种算法都要高，但由于在高维因果网络结构下，IGCI 错误地添加了很多边，准确率相对较低。同样，加性噪声模型(ANM)算法在高维数据情况下三个评分参数比 CLCI 和 TPCI 算法低。由于 CLCI 和 TPCI 算法采用了降维的方法，在维数增加的过程中，其保持了较好的稳定性，但是相对于利用 MI 的



(a) recall



(b) precision



(c)F_1

图 3　在不同维度数下四种算法的评分参数

TPCI,CLCI 利用 CDC 在大样本下的鲁棒性更好,时间复杂度更低,耗时更短。由图 3 也可以看出,CLCI 要优于其他三种算法。

4　结束语

与其他适用于高维数据的因果推断方法不同,文中结合基于信息论的 Copula 依赖系数,可以在高维和大样本数据下更为准确地检测出变量之间的关联关系,降低计算复杂度,再结合条件独立性测试对数据集进行无向结构学习,最后用 LSIR 算法对无向结构骨架进行每两点间的方向判断,得到最终的因果网络结构。文中采用虚拟网络进行的实验表明,利用 CDC 进行无向骨架的构造,耗时短,计算复杂度低于其他算法。当数据集维数较高、样本量大时,利用分治策略,该算法要大大优于其他因果推断算法。

参考文献:

[1] SCHAECHTLE U, STATHIS K, BROMURI S. Multi-dimensional causal discovery[C]//International joint conference on artificial intelligence. Beijing: AAAI Press, 2013: 1649-1655.

[2] SHIMIZU S. Non-Gaussian methods for causal structure learning[J]. Prevention Science, 2019, 20(3): 431-441.

[3] SPIRTES P, GLYMOUR C, SCHEINES R. Causation, prediction, and search[M]. Cambridge: MIT Press, 2000.

[4] TSAMARDINOS I, BROWN L E, ALIFERIS C F. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2006, 65(1): 31-78.

[5] CHEN Xuewen, ANANTHA G, LIN Xiaotong. Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm[J]. IEEE Transactions on Knowledge & Data Engineering, 2008, 20(5): 628-640.

[6] HOYER P O, JANZING D, MOOIJ J M, et al. Nonlinear causal discovery with additive noise models[C]//Proceedings of the 21st international conference on neural information processing systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008: 689-696.

[7] SHIMIZU S, HOYER P O, HYVÄRINEN A, et al. A linear non-Gaussian acyclic model for causal discovery[J]. Journal of Machine Learning Research, 2006, 7: 2003-2030.

[8] SHIMIZU S, INAZUMI T, SOGAWA Y, et al. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model[J]. Journal of Machine Learning Research, 2011, 12: 1225-1248.

[9] ZHANG K, HYVÄRINEN A. On the identifiability of the post-nonlinear causal model[C]//Proceedings of the 26th conference on uncertainty in artificial intelligence. [s. l.]: [s. n.], 2009: 647-655.

[10] LOH P L, BÜHLMANN P. High-dimensional learning of linear causal networks via inverse covariance estimation[J]. Journal of Machine Learning Research, 2013, 15: 3065-3105.

[11] HYVÄRINEN A, SMITH S M. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models[J]. Journal of Machine Learning Research, 2013, 14: 111-152.

[12] ZHANG K, PETERS J, JANZING D, et al. Kernel-based conditional independence test and application in causal discovery[C]//27th conference on uncertainty in artificial intelligence. [s. l.]: [s. n.], 2011: 804-813.

[13] YAMADA M, SUGIYAMA M, SESE J. Least-squares independence regression for non-linear causal inference under non-Gaussian noise[J]. Machine Learning, 2014, 96(3): 249-267.

[14] 曾千千,曾　安,潘　丹,等. 基于最大信息系数的贝叶斯网络结构学习算法[J]. 计算机工程, 2017, 43(8): 225-230.

[15] 张　浩,郝志峰,蔡瑞初,等. 基于互信息的适用于高维数据的因果推断算法[J]. 计算机应用研究, 2015, 32(2): 382-385.

[16] JIANG Hangjin, WU Qiongli. Robust dependence measure for detecting associations in large data set[J]. Acta Mathematica Scientia, 2018, 38(1): 57-72.