NOTEARS总结

□ 优化问题：组合无环约束转为光滑等式约束

$$\min_{G} \ score(\mathbf{X}, \mathbf{A}, \theta)$$

$$\text{s.t. } h(\mathrm{A}) = \operatorname{tr}\left(e^{\mathrm{A} \circ \mathrm{A}}\right) - d = 0$$

可行域非凸

□ 优化算法ALM：等式约束优化转为无约束优化

无约束目标函数：$L_c(\mathbf{A}, \theta, \lambda) = score(\mathbf{X}, \mathbf{A}, \theta) + \lambda h(\mathbf{A}) + \frac{c}{2}|h(\mathbf{A})|^2$

- **更新规则：** 超参数通常设为$\eta = 10$，$\gamma = \frac{1}{4}$，$\tau = 0.3$

$$\mathbf{A}_k, \theta_k = \arg\min_{\mathbf{A}, \theta} L_c(\mathbf{A}, \theta, \lambda)$$

随机优化L-BFGS-B求解，计算很耗时

拉格朗日乘数：$\lambda_{k+1} = \lambda_k + c_k h(\mathbf{A}_k)$

惩罚项参数：$c_{k+1} = \begin{cases} \eta c_k, & \text{if } |h(\mathbf{A}_k)| > \gamma |h(\mathbf{A}_{k-1})| \\ c_k, & \text{otherwise} \end{cases}$

$c \rightarrow \infty$确保无环性

优化停止准则：$h(\mathbf{A}_k) < \epsilon \in \{1e^{-6}, 1e^{-8}, 1e^{-10}\}$，不能保证输出DAG，需要设置阈值$\tau$，需要大量调参。

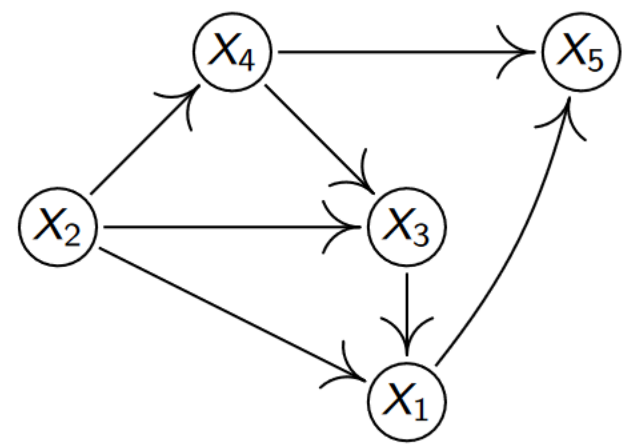
[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

因果发现-背景

□ **背景：** 现实很多统计问题，由于成本、风险和伦理问题，数据不能来自**RCT**，需要从**观察性数据**中获取因果关系。

observed iid data from $P(X_1, \ldots, X_5)$ $\overset{?}{\longrightarrow}$ causal model, e.g. DAG $\mathcal{G}$

X_1	X_2	X_3	X_4	X_5
3.4	−0.3	5.8	−2.1	2.2
1.7	−0.2	7.0	−1.2	0.4
−2.4	−0.1	4.3	−0.7	3.5
2.3	−0.3	5.5	−1.1	−4.4
3.5	−0.2	3.9	−0.9	−3.9
⋮	⋮	⋮	⋮	⋮



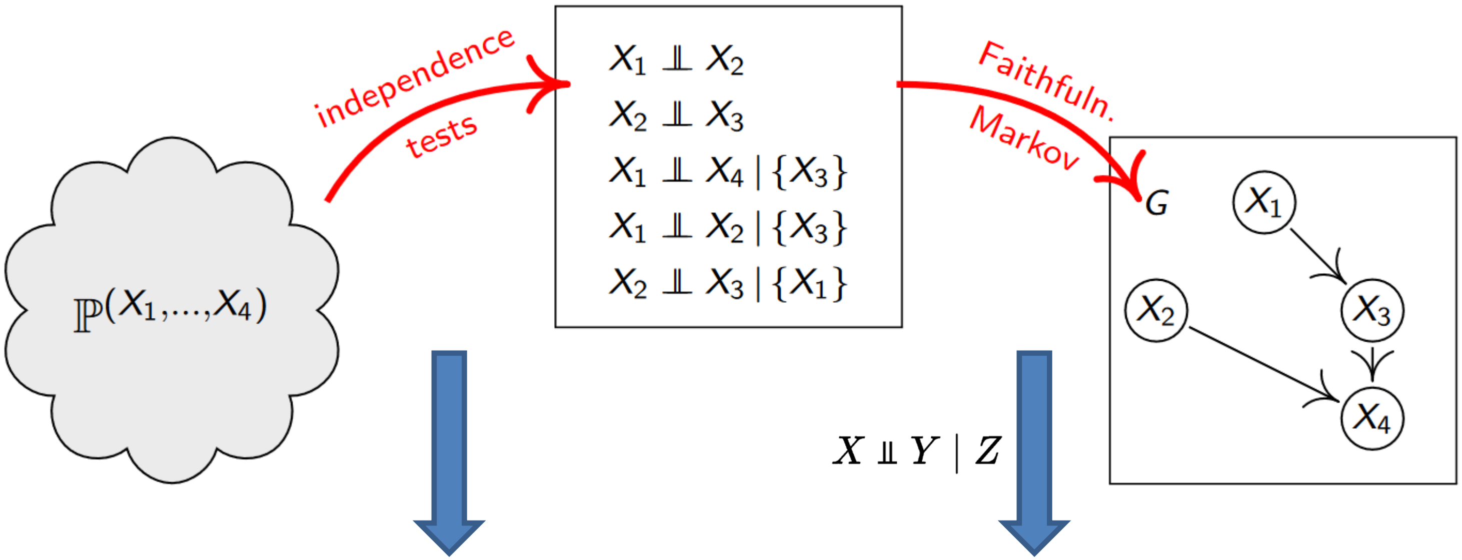
无法通过观测数据直接获取因果关系

[Jonas Peters, University of Copenhagen, Mini-course on Causality, Laboratory for Information & Decision Systems (LIDS) and Models, Inference & Algorithms of the Broad Institute, MIT, 2017]

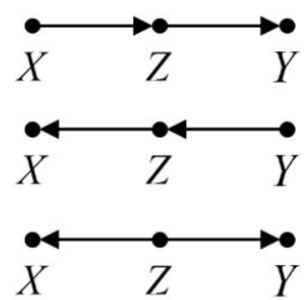
因果发现-研究现状

东南大学 SOUTHEAST UNIVERSITY

□ 基于条件独立性约束：PC



时间复杂度高，无法适用于高维复杂数据



马尔可夫等价类

无法解决MEC，无法适用于高维复杂数据

[Jonas Peters, University of Copenhagen, Mini-course on Causality, Laboratory for Information & Decision Systems (LIDS) and Models, Inference & Algorithms of the Broad Institute, MIT, 2017]

因果发现-研究现状

东南大学 SOUTHEAST UNIVERSITY

□ 基于因果函数的模型：

【定义】加性噪声模型：如果 X 和 Y 的关系可以用如下的一个函数加一个噪声的SCM形式来表示的话，那么我们称联合概率分布$P_{X,Y}$ 满足 X 到 Y 的加性噪声模型：$Y = f_Y(X) + N_Y, N_Y \perp\!\!\!\perp X$

基于加性噪声模型的可辨识性定义，判断 X 和 Y 之间因果关系的方法如下：

1. 在 X 上回归 Y ，即，用某种回归算法将 Y 表示为一个关于 X 的函数 f_Y ，加上某个噪声。
2. 测试 $Y - \hat{f}_Y(X)$ 是否与 X 独立。
3. 重复以上步骤但反过来在 Y 上回归 X 。
4. 如果这两种情形的测试结果，某一个独立，另一个不独立，那么独立的那个就是因果关系的方向。

依赖数据分布假设+因果充分性

[Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms[M]. The MIT Press, 2017.]

因果发现-研究现状

东南大学 SOUTHEAST UNIVERSITY

□ 基于因果函数的模型：3种假设

模型	函数 F	误差 E	
LiNGAM	$X = BX + E$	E is non-Gaussian	①
ANM	$Y = f(X) + E, f$ is nonlinear	$X \perp\!\!\!\perp E$	②
PNL	$Y = f_2(f_1(X) + E)$	$X \perp\!\!\!\perp E$	②
IGCI	$Y = f(X), f$ is nonlinear, $X \perp\!\!\!\perp f$	–	③
HCR	$X \to Y' \to Y, X \perp\!\!\!\perp f$	–	③

【定理1】考虑一个如下的结构因果模型 $X \to Y$：

$$Y = \alpha X + N_Y, N_Y \perp\!\!\!\perp X$$

那么，存在 $\beta \in \mathbb{R}$ 使得

$$X = \beta Y + N_X, N_X \perp\!\!\!\perp Y$$

成立，当且仅当 (X, N_Y) 的分布是高斯的。

【定理2】考虑一个如下的结构因果模型 $X \to Y$：

$Y = f(X) + N_Y, N_Y \perp\!\!\!\perp X$

$N_Y, X \sim \mathcal{N}$

那么只有当 f 是线性函数时才会有反向模型存在：

$$X = g(Y) + N_X, N_X \perp\!\!\!\perp Y$$
$$N_X, Y \sim \mathcal{N}$$

[Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms[M]. The MIT Press, 2017.]

因果发现-研究现状

□ 基于评分的方法：

➢ 传统评分方法问题描述：

$$\min_{\mathrm{G}} \quad Q(\mathrm{G})$$
$$\text{subject to} \quad \mathrm{G} \in \mathbb{D}$$

依赖评分的设计

NP-hard问题

a discrete score $Q : \mathbb{D} \to \mathbb{R}$ over the set of DAGs $\mathbb{D}$.

n个顶点的所有DAG数量： $a_n = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$

当 $n = 0, 1, 2, 3, \ldots$ ， $a_n = 1, 1, 3, 25, 543, 29281, 3781503, \ldots$

NP-hard问题：随着节点数增加，DAG数量呈超指数增长

[Robinson R W. Counting labeled acyclic digraphs[J]. New directions in the theory of graphs, 1973: 239-273.]

NOTEARS理论

□ 将$loss$函数转为定义在连续空间上：

$$\min_{\mathsf{G}} \quad Q(\mathsf{G}) \qquad \text{subject to} \quad \mathsf{G} \in \mathbb{D}$$

$$\Rightarrow$$

$$\min_{W \in \mathbb{R}^{d \times d}} \quad F(W) \qquad \text{subject to} \quad \mathsf{G}(W) \in \mathbb{D}.$$

$$W \in \mathbb{R}^{d \times d}$$

$$[\mathcal{A}(W)]_{ij} = 1 \iff w_{ij} \neq 0$$

考虑线性*SEM*：
$$\begin{cases} X = (X_1, \ldots, X_d)^T \\ W = [w_1 | \cdots | w_d] \\ X_j = w_j^T X + z_j \\ z = (z_1, \ldots, z_d) \end{cases}$$

$$\xrightarrow{\text{数据生成机制}}$$

$$\mathbf{X} = \mathbf{X}W + Z$$
$$\mathbf{X} \in \mathbb{R}^{n \times d}$$

目标函数： $F(W) = \ell(W; \mathbf{X}) + \lambda \|W\|_1 = \frac{1}{2n} \|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda \|W\|_1.$

$$\|W\|_1 = \|\operatorname{vec}(W)\|_1$$

DAG约束仍为组合约束，难以执行

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS理论

□ 将组合无环约束转为光滑等式约束，方便优化求解：

$$\min_{W \in \mathbb{R}^{d \times d}} \quad F(W) \qquad\qquad \min_{W \in \mathbb{R}^{d \times d}} \quad F(W)$$

$$\text{subject to} \quad \mathsf{G}(W) \in \mathbb{D}. \quad \Longleftrightarrow \quad \text{subject to} \quad h(W) = 0.$$

we would like a function $h : \mathbb{R}^{d \times d} \to \mathbb{R}$ that satisfies the following desiderata:

(a) $h(W) = 0$ if and only if W is acyclic (i.e. $\mathsf{G}(W) \in \mathbb{D}$);
(b) The values of h quantify the "DAG-ness" of the graph;
(c) h is smooth;
(d) h and its derivatives are easy to compute.

Theorem 1. *A matrix $W \in \mathbb{R}^{d \times d}$ is a DAG if and only if*

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0, \tag{5}$$

where $\circ$ is the Hadamard product and e^A is the matrix exponential of A. Moreover, $h(W)$ has a simple gradient

$$\nabla h(W) = \left(e^{W \circ W}\right)^T \circ 2W, \tag{6}$$

and satisfies all of the desiderata (a)-(d).

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

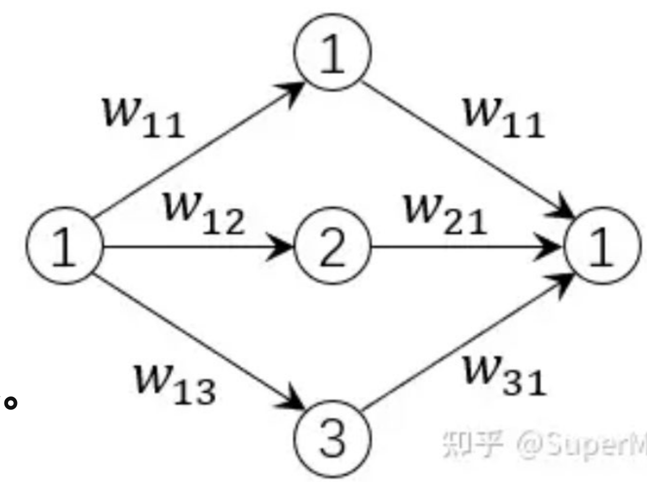
NOTEARS理论

口 定理证明： 先考虑3个节点的例子

$$\begin{bmatrix} w'_{11} & w'_{12} & w'_{13} \\ w'_{21} & w'_{22} & w'_{23} \\ w'_{31} & w'_{32} & w'_{33} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix}$$

通过2步，从1到1的路线权重总和：

$$w'_{11} = w_{11} \cdot w_{11} + w_{12} \cdot w_{21} + w_{13} \cdot w_{31}$$



假设权重均非负，推广得：

$tr(W^k) = 0$ 表示通过k步，所有节点均不成环。

真正DAG要求无论走多少步，均不成环。

DAG约束 ⟺ $\mathrm{tr}(W) + \mathrm{tr}(W^2) + \mathrm{tr}(W^3) + \ldots + \mathrm{tr}(W^k) + \ldots = 0$

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.] [Some pictures from zhihu.]

NOTEARS理论

东南大学 SOUTHEAST UNIVERSITY

DAG约束 $\Longleftrightarrow$ $\mathrm{tr}(W)+\mathrm{tr}\left(W^2\right)+\mathrm{tr}\left(W^3\right)+\ldots+\mathrm{tr}\left(W^k\right)+\ldots=0$

$$\Updownarrow$$

$$\forall \boldsymbol{i}, \boldsymbol{k}:\ \left(W^k\right)_{ii}=0$$

$$\Updownarrow$$

$$\sum_{k=1}^{\infty} \boxed{\sum_{i=1}^{d}\left(W^k\right)_{ii}} / k! = \mathrm{tr}\left(e^W\right)-d=0$$

$= \boldsymbol{tr(W^k)}$：**量化了图的无环性**

$$\mathrm{tr}\left(e^W\right)=\mathrm{tr}(I)+\mathrm{tr}(W)+\frac{1}{2!}\mathrm{tr}\left(W^2\right)+\cdots$$

为保证W中元素全非负，使用 $W \circ W$ 替换 W，最终得：

Theorem 1. *A matrix $W \in \mathbb{R}^{d \times d}$ is a DAG if and only if*

$$h(W)=\mathrm{tr}\left(e^{W \circ W}\right)-d=0,$$

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS优化

□ 优化方法：

$$\min_{W \in \mathbb{R}^{d\times d}} \quad F(W) \qquad\qquad \min_{W \in \mathbb{R}^{d\times d}} \quad F(W)$$

$$\text{subject to} \quad \mathsf{G}(W) \in \mathbb{D}. \quad \Longleftrightarrow \quad \text{subject to} \quad h(W) = 0.$$

增广拉格朗日法 $\Downarrow$ $\quad \boldsymbol{w} = \mathrm{vec}(W) \in \mathbb{R}^p$

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} f(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1,$$

$$\text{where} \quad f(\boldsymbol{w}) = \ell(W; \mathbf{X}) + \frac{\rho}{2}|h(W)|^2 + \alpha h(W)$$

通过增广拉格朗日法求解等式约束优化ECP

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS优化实现

Algorithm 1 NOTEARS algorithm

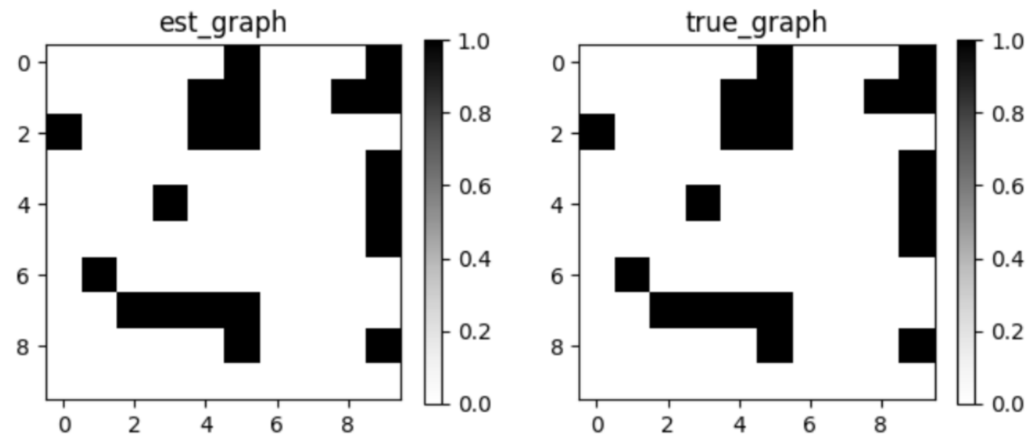
1. Input: Initial guess (W_0, α_0), progress rate $c \in (0, 1)$, tolerance $\epsilon > 0$, threshold $\omega > 0$.
2. For $t = 0, 1, 2, \ldots$:
   (a) Solve primal $W_{t+1} \leftarrow \arg\min_W L^\rho(W, \alpha_t)$ with ρ such that $h(W_{t+1}) < ch(W_t)$.
   (b) Dual ascent $\alpha_{t+1} \leftarrow \alpha_t + \rho h(W_{t+1})$.
   (c) If $h(W_{t+1}) < \epsilon$, set $\widetilde{W}_{\mathrm{ECP}} = W_{t+1}$ and break.
3. Return the thresholded matrix $\widehat{W} := \widetilde{W}_{\mathrm{ECP}} \circ 1(|\widetilde{W}_{\mathrm{ECP}}| > \omega)$.

```python
n, d = X.shape      # n=100, d=20
w_est, rho, alpha, h = np.zeros(2 * d * d), 1.0, 0.0, np.inf  # double w_est into (w_pos, w_neg)
bnds = [(0, 0) if i == j else (0, None) for _ in range(2) for i in range(d) for j in range(d)]  #
if loss_type == 'l2':
    X = X - np.mean(X, axis=0, keepdims=True)
for _ in range(max_iter):       #100次
    w_new, h_new = None, None
    while rho < rho_max:        #rho_max=1e+16
        sol = sopt.minimize(_func, w_est, method='L-BFGS-B', jac=True, bounds=bnds)
        w_new = sol.x                   #取得最优时的w_new(800,)
        h_new, _ = _h(_adj(w_new))
        if h_new > 0.25 * h:            #如果h变化不大，则rho增大，否则跳出循环
            rho *= 10
        else:
            break
    w_est, h = w_new, h_new
    alpha += rho * h
    if h <= h_tol or rho >= rho_max:
        break
W_est = _adj(w_est)
W_est[np.abs(W_est) < w_threshold] = 0
return W_est
```

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS实验

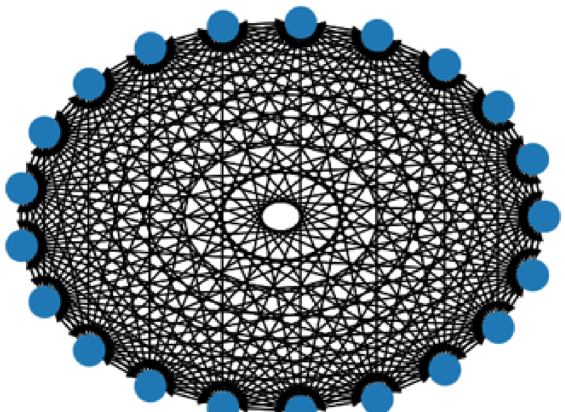
```python
# simulate data for notears
weighted_random_dag = DAG.erdos_renyi(n_nodes=10, n_edges=20, weight_range=(0.5, 2.0), seed=1)
dataset = IIDSimulation(W=weighted_random_dag, n=2000, method='linear', sem_type='gauss')
true_dag, X = dataset.B, dataset.X

# notears learn
nt = Notears()
nt.learn(X)

# plot est_dag and true_dag
GraphDAG(nt.causal_matrix, true_dag)

# calculate accuracy
met = MetricsDAG(nt.causal_matrix, true_dag)
print(met.metrics)
```



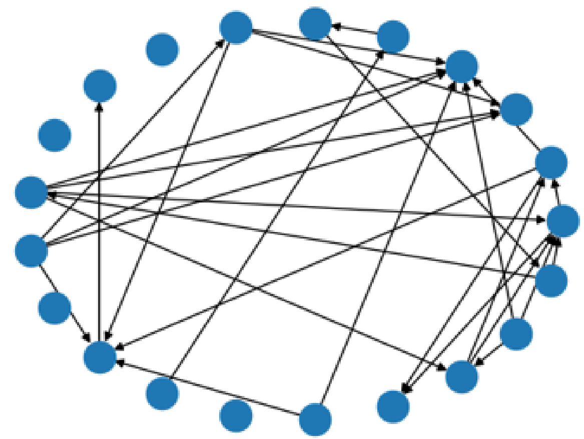
[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS实验

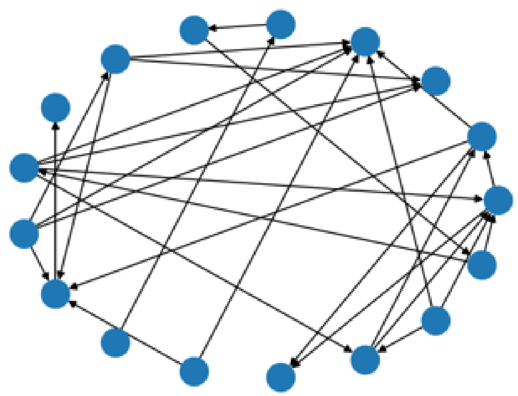
□ 通过增加阈值，可显著提升因果发现准确率：



不加阈值



加阈值



删除孤立节点后

通过增加阈值，可显著提升因果发现准确率

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]

NOTEARS优缺点

☐ 优点：

- 提出基于连续优化从数据中学习DAG，取代了原先困难的离散优化，进行平滑的全局搜索，而不是组合的局部搜索。

☐ 缺点：

- NOTEARS对DAG的搜索利用标准数值求解器进行（黑盒优化过程，求得局部最优解），从DICD和ReScore可看出优化过程不稳定。
- 只能处理变量间线性关系。
- 最终二分类阈值的选取需要经验。

[Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.]