PC ALGORITHM 实现修正

理想情况（知道所有条件独立关系）

Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm, Markus Kalisch, Peter Bu¨hlmann. 2007

Algorithm 1 The PC_{pop}-algorithm

1: **INPUT:** Vertex Set V, Conditional Independence Information
2: **OUTPUT:** Estimated skeleton C, separation sets S (only needed when directing the skeleton afterwards)
3: Form the complete undirected graph $\tilde{C}$ on the vertex set V.
4: $\ell = -1$; $\quad C = \tilde{C}$
5: **repeat**
6:     $\ell = \ell + 1$
7:     **repeat**
8:         Select a (new) ordered pair of nodes i, j that are adjacent in C such that $|adj(C,i) \setminus \{j\}| \geq \ell$
9:         **repeat**
10:             Choose (new) $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$.
11:             **if** i and j are conditionally independent given $\mathbf{k}$ **then**
12:                 Delete edge i, j
13:                 Denote this new graph by C
14:                 Save $\mathbf{k}$ in $S(i,j)$ and $S(j,i)$
15:             **end if**
16:         **until** edge i, j is deleted or all $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been chosen
17:     **until** all ordered pairs of adjacent variables i and j such that $|adj(C,i) \setminus \{j\}| \geq \ell$ and $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been tested for conditional independence
18: **until** for each ordered pair of adjacent nodes i,j: $|adj(C,i) \setminus \{j\}| < \ell$.

line 11: 需要条件独立关系

现有数据无法进行条件独立性检测

偏相关系数

校正其它变量后某一变量与另一变量的相关关系，校正的意思可以理解为假定其它变量都取值为均数

服从高斯分布的随机变量，条件独立性与偏相关系数为0等价：

假设随机变量 X 服从多元高斯分布，对于$i \neq j \in (1, \ldots, p)$，$k \subseteq (1, \ldots, p)/(i, j)$，用 $\rho_{i,j|k}$ 表示 $X(i)$ 和 $X(j)$ 与 $X^{(r)}(r \in k)$ 之间的偏相关系数。当且仅当 $X(i)$ 和 $X(j)$ 条件独立与 $X^{(r)}(r \in k)$ 时，$\rho_{i,j|k} = 0$。

∴ 条件独立性可由偏相关估计出来，条件独立性检验转偏相关系数检验

任意两个变量i, j的h（排除其他h个变量的影响后，$h <= k - 2$）阶样本偏相关系数：

$$\rho_{i,j|\mathbf{k}} = \frac{\rho_{i,j|\mathbf{k}\backslash h} - \rho_{i,h|\mathbf{k}\backslash h}\rho_{j,h|\mathbf{k}\backslash h}}{\sqrt{(1-\rho^2_{i,h|\mathbf{k}\backslash h})(1-\rho^2_{j,h|\mathbf{k}\backslash h})}}$$

Z检验（$\rho \neq 0$时的显著性检验）

$\rho \neq 0$时不是正态分布，不能进行 t 检验。将 ρ 进行 Fisher Z 转换，转换后可以认为是正态分布。

Fisher's z-transform:

$$Z(i, j|\mathbf{k}) = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{i,j|\mathbf{k}}}{1-\hat{\rho}_{i,j|\mathbf{k}}}\right)$$

零假设：$H_0(i, j|k) : \rho_{i,j|k} \neq 0$

对立假设：$H_1(i, j|k) : \rho_{i,j|k} = 0$

当$\sqrt{n - |k| - 3}|Z(i, j|k) > \Phi^{-1}(1 - \alpha/2)$，$H_0$成立

∴ 用 $\sqrt{n - |k| - 3}|Z(i, j|k) <= \Phi^{-1}(1 - \alpha/2)$ 替换 PC-Algorithm 中的 " 如果 i, j 被 k $d - separation$"

paper: Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely large population, Fisher, R.A., 1915

R语言实现

`zStat(x, y, S, C)`：计算并返回$\sqrt{n-|k|-3}|Z(i,j|k)$的值

`pcorOrder(i, j, k, C)`：计算并返回 i 和 j 与 k 的偏相关系数

`condIndFisherZ(x, y, S, C)`：计算$\sqrt{n-|k|-3}|Z(i,j|k)$，返回它是否<= cutoff

`gaussCItest(x, y, S, suffStat)`：计算并返回$\Phi^{-1}(1-\alpha/2)$

原版PC Algorithm没有的步骤

An algorithm for fast recovery of sparse causal graphs, Peter Spirtes, Clark N. Glymour., 1990

A.) Form the complete undirected graph C on the vertex set V.

B.)

n = 0.

repeat

For each pair of variables a, b adjacent in C, if Acab n Ucab has cardinality greater than or equal to n and a, b are independent conditional on any subsets of Acab n Ucab of cardinality n, delete a-b from C.

n = n + 1.

until for each pair of adjacent vertices a, b, Acab n Ucab is of cardinality less than n.

C.) Let F be the graph resulting from step B. For each triple of vertices a, b, c such that the pair a, b and the pair b,c are each adjacent in F but the pair a, c are not adjacent in F, orient a - b - c as a -> b <- c if and only if a and c are dependent on every subset of Apac n Upac containing b. Output all graphs consistent with these orientations.

得到骨架（无向图）。

一些定义：

骨架：把有向图 G 的有向边变成无向边。

PDAG：设 $G=(V,E)$ 是一个图，若边集 E 中包含有向边和无向边，则称P是一个部分有向图。若部分有向图 P 中不存在有向圈，则称 P 是一个部分有向无环图(PDAG)

马尔科夫等价：贝叶斯网络 $<G_1,P_1>$ 和 $<G_2,P_2>$马尔科夫等价，当且仅当 G_1 和 G_2 具有相同的框架和V结构

有向无环图 $G=(V,E)$，任意有向边 $V_i \to V_j \in E$，若存在图 $G'=(V,E')$ 与 G 等价，且 $V_j \to V_i \in E'$，则称有向边 $V_i \to V_j$ 在 G 中是可逆的，否则是不可逆的.。

同理，对任意无向边 $V_i \to V_j \in E$，若存在 $G_1=(V,E_1)$、 $G_2=(V,E_2)$ 均与 G 等价，且 $V_i \to V_j \in E_1$、$V_j \to V_i \in E_2$，则 称 无 向边 $V_i \to V_j$ 在 G 中是可逆的，否则是不可逆的

CPDAG：设 $G=(V,E)$ 是一个部分有向无环图，若 E 中的有向边都是不可逆的，并且 E 中的无向边都是可逆的，则称 G 是一个完全部分有向无环图(CPDAG)

任意 BN 的马尔科夫等价类都存在唯一的 CPDAG 与之等价，因此，CPDAG可作为贝叶斯网络等价类的图形化表示

将骨架扩展为等价类的CPDAG：

Causal inference and causal explanation with background knowledge, Meek., 1995

Algorithm 2 Extending the skeleton to a CPDAG

INPUT: Skeleton G_{skel}, separation sets S
OUTPUT: CPDAG G
for all pairs of nonadjacent variables i, j with common neighbour k **do**
  if $k \notin S(i,j)$ **then**
    Replace $i - k - j$ in G_{skel} by $i \to k \leftarrow j$
  end if
end for
In the resulting PDAG, try to orient as many undirected edges as possible by repeated application of the following three rules:
R1 Orient $j - k$ into $j \to k$ whenever there is an arrow $i \to j$ such that i and k are nonadjacent.
R2 Orient $i - j$ into $i \to j$ whenever there is a chain $i \to k \to j$.
R3 Orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to j$ and $i - l \to j$ such that k and l are nonadjacent.
R4 Orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to l$ and $k \to l \to j$ such that k and l are nonadjacent.

TPDA

一种看起来挺靠谱的算法

Learning Belief Networks from Data: An Information Theory Based Approach