中国科学: 数学 2018 年 第 48 卷 第 12 期: 1753 ~ 1778
SCIENTIA SINICA Mathematica
综 述

因果推断的统计方法

苗旺[1], 刘春辰[2], 耿直[3*]

1. 北京大学光华管理学院商务统计与经济计量系, 北京 100871;
2. 日本电气 (NEC) 中国研究院, 北京 100600;
3. 北京大学数学科学学院, 北京 100871
E-mail: mwfy@gsm.pku.edu.cn, liu_chunchen@nec.cn, zhigeng@pku.edu.cn

收稿日期: 2018-03-12; 接受日期: 2018-09-03; 网络出版日期: 2018-11-20; * 通信作者
国家高技术研究发展计划 (批准号: 2015AA020507)、国家重点基础研究发展计划 (批准号: 2015CB856000) 和国家自然科学基金 (批准号: 11331011, 11771028 和 91630314) 资助项目

摘要 探索事物之间的因果关系和因果作用是很多科学研究的重要目的. 因果推断的统计方法是利用试验性研究和观察性研究得到的数据, 评价变量之间的因果作用和挖掘多个变量之间的因果关系. 本文将介绍因果作用和因果关系的形式化定义, 以及因果推断的两个主要统计模型: 潜在结果模型和因果网络模型. 本文将探讨因果作用的可识别性和因果网络的结构学习, 综述有关因果推断的若干研究问题和动态.

关键词 因果作用 因果网络 混杂因素 潜在结果模型 替代指标 因果推断 有向无环图

MSC (2010) 主题分类 62A01, 68T30

1 引言

探求事物之间的因果关系是哲学、自然科学和社会科学等众多研究所追求的终极目标. 古希腊哲学家 Democritus (约公元前 400 年) 认为: 发现一个因果关系胜过做国王. Bacon (培根) (1561–1626 年) 提出 "知识就是力量", 并认为 "真正的知识是根据因果关系得到的知识". 探索并发现因果关系及其方法论的研究, 伴随人类社会的发展而不断精深, 成为经久不衰的挑战. 如何根据观测和试验探究事物之间的因果关系, 这个问题自东西方古代哲学到现代科学, 已经困惑了人类数千年. 因果关系和相关关系是两个不同的概念. 即使两个变量有相关关系也可能没有因果关系; 相反地, 它们没有相关关系也可能有因果关系. 19 世纪末, 统计学家提出了各种相关关系的形式化度量, 最具有代表性的是 Pearson 相关系数. 一个多世纪以来, 统计学中关于相关关系的研究取得了令人瞩目的成就, 而关于因果关系的研究则进展缓慢. 正如文献 [1] 指出的那样, 涉及因果推断的问题自始就缠住了统计学前进的步伐. 2011 年, 图灵奖获得者 Pearl [2] 认为: 在过去的一个世纪中, 许多科学发现被推迟是由于缺少描述因果的数学语言. 近年来, 探索因果关系的研究越来越激励着统计学者和计算机学者, 在统

英文引用格式: Miao W, Liu C C, Geng Z. Statistical approaches for causal inference (in Chinese). Sci Sin Math, 2018, 48: 1753–1778, doi: 10.1360/N012018-00055

计领域中对因果推断的兴趣正在复兴 (参见文献 [3]). 2008 年以来, 在国际机器学习会议和期刊中组织了多次因果挑战的 Workshops [4] 和专题论文 [5]. 在大数据时代, 不仅要紧盯事物之间的相关关系, 更应该挖掘蕴藏在大数据中的因果关系 (defense advanced research projects agency, 简称 DARPA, Big Mechanism Program, https://www.darpa.mil/program/big-mechanism). 例如, 某超市发现顾客买尿布与买啤酒有很大的相关性, 实际上二者没有因果关系, 而出现这个相关性的原因是在家照顾婴儿的主妇常让丈夫去超市买尿布, 这些丈夫买尿布的同时会顺便买啤酒. 如果这些主妇改为自己在网上买尿布的话, 将不能期望她们的丈夫同时会买啤酒.

在科学研究的各个领域都存在因果推断的挑战问题. 在经济学中, 诺贝尔经济学奖获得者 Heckman [6] 提出了政策评价中的三个因果推断的挑战问题:

(1) 评价历史上出现的干预对结果的影响;

(2) 预测在一个环境中曾执行过的干预在其他环境中的影响;

(3) 预测历史上从没有经历过的干预在各种环境中的影响.

从时间序列预测的角度, 文献 [7] 提出了因果关系的定义: 如果利用 X 能更好地预测 Y, 那么 X 是 Y 的原因. 这个定义不能描述真正的因果关系, 实际上是相关关系. 例如, 看到闪电可以预测雷声, 而闪电却不是雷声的真正原因. 相关关系是 "预测" 的基础, 而因果关系是 "决策" 的基础.

目前, 因果推断采用的两个主要模型是潜在结果模型 [8, 9] 和因果网络模型 [10, 11]. 潜在结果模型给出了因果作用的数学定义. 该模型主要用在原因和结果变量已知的前提下, 定量评价原因变量对结果变量的因果作用. 因果网络模型是描述数据产生机制和外部干预的形式化语言. 因果网络是将 Bayes 网络加上外部干预, 用来定义外部干预的因果作用和描述多个变量之间的因果关系. 利用因果网络不仅能定量评价因果作用, 还能定性确定混杂因素, 用于从数据挖掘因果关系.

本文第 2 节介绍因果推断的潜在结果模型、因果作用的可识别性和随机化试验; 第 3 节介绍观察性研究和混杂因素, 以及混杂因素完全观测时的因果推断方法; 第 4 节介绍存在未观测的混杂因素时因果作用的估计方法; 第 5 节介绍替代指标悖论和确定替代指标的准则; 第 6 节介绍因果网络模型和结构学习的算法; 第 7 节展望因果推断在现代大数据研究中的前景.

2 潜在结果模型

2.1 潜在结果和因果作用的定义

两个变量 X 和 Y 的相关系数可以用 X 和 Y 的联合分布的函数表示. 例如, Pearson 相关系数表示为 $\rho(X,Y)=\sigma_{xy}/(\sigma_x\sigma_y)$, 其中 σ_{xy} 是 X 和 Y 的协方差, σ_x 和 σ_y 分别是 X 和 Y 的标准差. 但是, 如何形式化表示两个变量 X 和 Y 的因果作用的度量呢? 用观测变量 X 和 Y 的联合分布的函数怎么也不能清楚地定义因果作用.

统计学家利用潜在结果给出了因果作用的形式化定义. 利用潜在结果模型, 文献 [8] 针对试验性研究 (experimental studies) 给出了因果作用的数学定义, 文献 [9] 将这一定义推广到观察性研究 (observational studies). 潜在结果模型通常需要假定个体处理值稳定 (stable unit treatment value assumption, SUTVA): 个体 i 的潜在结果不受其他个体的处理的影响, 并且对每个个体和每一种处理只有一个潜在结果, 详细讨论参见文献 [12]. 考虑一个二值处理或暴露变量 (treatment/exposure), $X=1$ 代表处理组, 0 代表对照组. 在接受处理分配 $X=x$ 后的结果变量为 Y_x, 表示假如接受处理 $X=x$ 后的结果, 称为潜在结果 (potential outcome). 对每个个体, 实际观察的结果变量 Y 可以用潜在结果 (Y_1,Y_0) 表

示为 $Y = XY_1 + (1-X)Y_0$. SUTVA 假定意味着张三的病是否被治好不受李四是否吃药的影响. 这个假定在很多实际问题中不成立, 例如, 李四获奖, 也许会影响其同事张三的工作积极性; 周围的朋友打不打流感预防针也许会影响自己得不得流感. 这个假定也许是潜在结果模型的重要缺陷之一. 目前有一些学者试图利用社会网络方法弥补这个缺陷 (参见文献 [13–18]).

因果作用定义为相同个体的潜在结果之差 (参见文献 [8, 9, 19]). 个体 i 的因果作用 (individual causal effect, ICE) 定义为 $\mathrm{ICE}(i) = Y_1(i) - Y_0(i)$. 尽管潜在结果模型清楚地定义了个体因果作用, 但是, 正如 Heraclitus (古希腊哲学家) 所指出的那样: 你不可能两次踏入相同的河. 对于每个个体 i, 通常不可能既观测到 $Y_1(i)$, 又观测到 $Y_0(i)$. 因此, 个体因果作用通常是不能从观测数据推断的. 尽管如此, 有一些学者试图探讨个体因果作用的统计推断方法, 这一般需要较强的模型假定. 目前, 个体化治疗和精准治疗也是试图推断个体因果作用或同质个体人群的因果作用 (参见文献 [20–23]). 由于对每一个个体 i, $Y_1(i)$ 和 $Y_0(i)$ 不能同时观测到, 因此, 因果推断还可以看作是一个缺失数据的问题.

统计学关心的是总体的特征, 利用潜在结果, 还可以定义总体的平均因果作用.

定义 2.1 总体的平均因果作用 (average causal effect, ACE) 定义为个体因果作用的期望:

$$\mathrm{ACE} = \mathrm{E}(\mathrm{ICE}) = \mathrm{E}(Y_1 - Y_0) = \mathrm{E}(Y_1) - \mathrm{E}(Y_0).$$

平均因果作用定义为假若所有个体都接受处理 $X = 1$ 的平均结果 $\mathrm{E}(Y_1)$ 与假若所有个体都接受对照 $X = 0$ 的平均结果 $\mathrm{E}(Y_0)$ 之差. 在实际中不可能让所有的个体都接受处理 $X = 1$, 再接受对照 $X = 0$; 即使这样做, 对同一个体 i, 先接受处理 $X = x$ 的潜在结果 $Y_x(i)$ 与后接受处理 $X = x$ 的潜在结果 $Y'_x(i)$ 可能也不一致.

进一步, 人们可能关心某个子总体的平均因果作用. 例如, 某药物对男性和女性分别的疗效.

定义 2.2 令 V 为协变量. 定义 $V = v$ 子总体的平均因果作用为 $\mathrm{E}(Y_1 - Y_0 \mid V = v)$.

另外, 人们常常关心处理组的因果作用. 例如, 流行病学家并不关心吸烟对整个人群的因果作用, 而只关心吸烟对吸烟人群的因果作用.

定义 2.3 处理组的平均因果作用定义为 $\mathrm{E}(Y_1 - Y_0 \mid X = 1)$.

称平均因果作用 $\mathrm{ACE} = \mathrm{E}(Y_1 - Y_0)$ 为可识别的, 如果 ACE 可以由观测变量的分布 $\mathrm{pr}(X, Y, V)$ 唯一确定. 如果 ACE 不可识别, 则意味着至少存在两个不相等的 $\mathrm{ACE} \neq \mathrm{ACE}'$ 满足观测到的数据. 可识别性往往是因果推断中最棘手的问题. 为了得到因果作用的可识别性, 通常需要有额外的假定. 随机化试验是识别因果作用最有效的方法.

2.2 随机化试验

统计学家 Fisher 给出了识别平均因果作用的方法: 随机化试验设计. 随机化处理分配 X 给个体 i, 例如, 抛硬币确定个体 i 的处理 X, 与潜在结果及协变量的取值无关, 可以保证潜在结果 (Y_1, Y_0) 与处理分配 X 独立, 即在随机化分配下, 有 $(Y_1, Y_0) \perp\!\!\!\perp X$, 进而,

$$\mathrm{E}(Y_x) = \mathrm{E}(Y \mid X = x), \quad \mathrm{ACE} = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0).$$

在随机化分配下, 平均因果作用表示为观测到的结果变量 Y 在处理组 $X = 1$ 与对照组 $X = 0$ 中期望之差, 不再含有潜在结果变量 Y_1 和 Y_0, 因此, 它是可识别的. 通过分别估计 $\mathrm{E}(Y \mid X = 1)$ 和 $\mathrm{E}(Y \mid X = 0)$, 传统的统计推断方法可以用来推断平均因果作用. 例如, 用 t- 检验研究平均因果作用是否为零. 随机化试验的理论、设计和实践, 参见文献 [24].

在实际研究中, 随机化试验往往不具备可操作性. 例如, 研究吸烟对肺癌的作用, 不能随机化分配一个人吸烟或不吸. 在实际中经常面临的其他问题, 如代价昂贵和个体不依从 (noncompliance) 等也都限制了随机化试验的应用. 在以下两节, 我们介绍利用观察性研究推断因果作用的方法. 这些方法使用观察性研究得到的数据, 通过引入处理分配可忽略性假定, 或者辅助变量, 如工具变量 (instrumental variable) 和阴性对照变量 (negative control variable) 来推断因果作用.

3 观察性研究和可忽略性

3.1 观察性研究和混杂因素

文献 [25] 定义观察性研究 (observational studies) 为根据经验观察推断因果作用的研究, 但不能采用有控制的试验, 也不能随机地分配处理. 观察性研究不再满足随机分配的条件, 即 $(Y_1, Y_0) \not\perp\!\!\!\perp X$. 在这种情形下, 如果忽略协变量, 仅用随机化试验推断方法估计因果作用就会导致偏差. 这个偏差甚至会造成悖论, 如 Yule-Simpson 悖论 [26, 27]. 解释这个悖论的一个很好的例子是 Bickel 等 [28] 关于加利福尼亚大学伯克莱分校的研究生入学中是否存在性别歧视的研究. 他们在观测数据中发现男生录取比例高于女生, 但是, 根据学生申请专业分层后, 发现女生的录取率略高于男生. 可见, 如果没有记录学生的报考专业数据, 或者不对此做调整, 就会产生完全错误的结论. 在这个例子中, 申请专业是研究性别对录取率影响的关键因素, 称为混杂因素, 忽略这样的因素会导致因果作用估计的偏差.

判断和确定哪些变量或因素是混杂因素的问题是因果推断中最基本和关键的问题. 判别混杂因素的准则大致分为两大类: 可压缩性准则和可比较性准则. 可压缩性准则根据相关关系的度量定义混杂因素. 如果相关关系的度量受第三个变量的影响, 那么该变量为混杂因素. 文献 [29] 介绍了混杂因素的可压缩性准则.

可比较性准则是基于潜在结果模型来定义混杂偏倚和混杂因素. 如果暴露总体的潜在结果 Y_1 和 Y_0 的分布分别与非暴露总体的潜在结果的分布相同 (即 $\mathrm{pr}(Y_1 \mid X = 1) = \mathrm{pr}(Y_1 \mid X = 0)$ 和 $\mathrm{pr}(Y_0 \mid X = 1) = \mathrm{pr}(Y_0 \mid X = 0)$), 则称暴露总体与非暴露总体是可比较的 (或称可交换的), 也称为无混杂偏倚. 在这种情形下, 暴露对结果的平均因果作用 $\mathrm{ACE} = \mathrm{E}(Y_1) - \mathrm{E}(Y_0)$ 等于暴露组与非暴露组观测结果的期望之差 $\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$. 文献 [30] 利用很多流行病学的案例归纳出判断混杂因素的准则: 混杂因素 V 必须同时满足下面两个条件:

(1) V 是一个独立的风险因素;

(2) V 在暴露总体与在非暴露总体中的分布不同.

文献 [31, 32] 将混杂因素叙述为, 当控制和调整某个变量使混杂偏倚减小时, 该变量为混杂因素. 因果推断的关键问题是, 如何推断潜在结果的条件期望 $\mathrm{E}(Y_{1-x} \mid X = x)$. 这个期望是反事实的, 因为对 $X = x$ 的个体, 潜在结果 Y_{1-x} 永远观测不到. 一种推断反事实的潜在结果期望的方法称为标准化方法. 标准化是指, 利用观测的协变量 V 对 $X = 1 - x$ 组的观测结果 Y 的期望进行调整 $\theta_{1-x} = \mathrm{E}[\mathrm{E}(Y \mid X = 1 - x, V)]$. 据此, 文献 [33] 给出了混杂因素的形式定义: 如果利用变量 V 的标准化能减少混杂偏倚, 即

$$|\mathrm{E}(Y_{1-x} \mid X = x) - \theta_{1-x}| < |\mathrm{E}(Y_{1-x} \mid X = x) - \mathrm{E}(Y \mid X = 1 - x)|,$$

则称变量 V 为混杂因素. 根据这个混杂因素的形式化定义, 可以证明文献 [30] 的准则 (形式化为 $V \not\perp\!\!\!\perp Y \mid X$ 和 $V \not\perp\!\!\!\perp X$) 是混杂因素的必要条件, 但不是充分条件. 也就是说, 文献 [30] 的准则只能

判断哪些变量不是混杂因素, 但不能判断哪些是混杂因素.

3.2 可忽略性假定

文献 [34] 提出的处理分配机制的可忽略性假定是观察性研究中判断混杂因素和推断因果作用最重要的假定. 本文的第 3.2–3.5 小节都假定如下的可忽略性成立.

假定 3.1 (处理分配机制的可忽略性 (ignorability of treatment assignment mechanism), 简称可忽略性) 令 V 表示观测的协变量, 如果满足 (i) $(Y_1, Y_0) \perp\!\!\!\perp X \mid V$ 和 (ii) $0 < \mathrm{pr}(X = 1 \mid V) < 1$, 那么称处理分配机制是可忽略的.

可忽略性假定中的条件 (i) 相当于在 V 的每一层做了随机化分配, 那么, 在 V 的每一层, 平均因果作用是可识别的, 进而对 V 求期望可以得到总体的平均因果作用. 条件 (ii) 要求在 V 的每一层里接受处理或对照的概率大于 0, 这是为了保证在每一层都能得到该层平均因果作用的相合估计. 在可忽略性假定下, 平均因果作用可通过以下公式识别:

$$\mathrm{E}(Y_1 - Y_0) = \mathrm{E}[\mathrm{E}(Y_1 - Y_0 \mid V)] = \mathrm{E}[\mathrm{E}(Y \mid X = 1, V) - \mathrm{E}(Y \mid X = 0, V)].$$

可忽略性假定解释了随机化试验和观察性研究的差别. 如果处理 X 没有随机分配, 而仅仅是可忽略性假定成立, 那么不对混杂因素进行调整, 就会导致混杂偏倚,

$$\begin{aligned}\{\mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)\} - \mathrm{ACE} = &\int_V \mathrm{E}(Y \mid X = 1, V)\{\mathrm{pr}(V \mid X = 1) - \mathrm{pr}(V)\}dV \\ &- \int_V \mathrm{E}(Y \mid X = 0, V)\{\mathrm{pr}(V \mid X = 0) - \mathrm{pr}(V)\}dV.\end{aligned}$$

当协变量 V 的分布在处理组和对照组不均衡时, 即 $\mathrm{pr}(V \mid X = 1) \neq \mathrm{pr}(V \mid X = 0)$ 时, 该混杂偏倚一般不为零, 因此在进行平均因果作用的统计推断时, 需要对协变量 V 做调整.

在可忽略性假定下, 多种统计推断方法可以用来估计因果作用. 例如, 当 V 是一个有 K 个水平的离散变量时, 可以先在 V 的每一层估计 $\mathrm{ACE}_k = \mathrm{E}(Y_1 - Y_0 \mid V = k)$, 然后估计

$$\mathrm{ACE} = \sum_{k=1}^{K} \mathrm{ACE}_k \mathrm{pr}(V = k).$$

但是, 当 V 是高维变量或连续型变量时, 按 V 的值将总体分层会导致每一层的样本太少, 增大估计的方差. 在这种情形下, 通常建立一些参数模型来估计因果作用.

3.3 倾向得分和匹配

为了消除协变量的分布在处理组与对照组之间的差异, 匹配 (matching) 方法经常用在观察性研究中. 匹配方法的目的是对每一个个体匹配一个具有相同或相近协变量取值的个体集合, 使得匹配得到的数据在处理组和对照组有相同的协变量分布, 然后根据匹配数据推断因果作用. 早期的匹配方法根据一个或几个协变量直接构造匹配集合 [35, 36]. 但是在很多应用中, 协变量维数较高, 难以决定根据哪些协变量构造匹配集合. 文献 [34] 提出了倾向得分匹配 (propensity score matching) 方法, 根据一个一维的倾向得分构造匹配集合, 目前已经是观察性研究中常使用的匹配方法.

定义 3.1 倾向得分定义为条件概率 $\pi(V) = \mathrm{pr}(X = 1 \mid V)$.

定理 3.1 [34] 如果给定协变量 V 时可忽略性成立, 即 $Y_x \perp\!\!\!\perp X \mid V$ 且 $0 < \mathrm{pr}(X = 1 \mid V) < 1$, 那么, 给定倾向得分 $\pi(V)$ 时可忽略性也成立, 即 $Y_x \perp\!\!\!\perp X \mid \pi(V)$ 且 $0 < \mathrm{pr}\{X = 1 \mid \pi(V)\} < 1$.

因此, 可以利用倾向得分分层或匹配进行因果推断, 从而避免了用高维协变量 V 进行分层或匹配的困难. 给定样本中个体 $i=1,\ldots,n$, 采用文献 [37] 中的做法, 个体 i 根据倾向得分得到的匹配集合定义为

$$J(i)=\left\{j=1,\ldots,n: X_j=1-X_i; \sum_{k:X_k=1-X_i}\delta\{|\pi(V_i)-\pi(V_k)|\leqslant|\pi(V_i)-\pi(V_j)|\}\leqslant M\right\},$$

其中 $\delta\{\cdot\}$ 是示性函数, 当括号中的条件满足时取值 1, 否则取值 0; M 为整数, 代表每一个个体的匹配数据的个数, 例如, $M=1$ 时得到 $1:1$ 匹配. 这个定义允许在构造匹配集合过程中放回已被使用的个体, 并用在其他个体的匹配集合中. 平均因果作用的匹配估计为

$$\widehat{\phi}_{\rm psm}=\frac{1}{n}\sum_{i=1}^{n}(2X_i-1)\left(Y_i-\frac{1}{M}\sum_{j\in J(i)}Y_j\right).$$

如果在实际中不知道真实的倾向得分, 可以根据数据预先估计, 然后用估计得到的倾向得分做匹配. 常用的估计倾向得分方法包括 logistic 回归和决策树等机器学习方法 [38].

匹配方法还可以用来估计处理组的平均因果作用

$$\widehat{\phi}_{\rm psm1}=\frac{1}{\sum_{i=1}^{n}X_i}\sum_{i=1}^{n}X_i\left(Y_i-\frac{1}{M}\sum_{j\in J(i)}Y_j\right).$$

在一定正则条件下, 可以证明匹配估计的相合性和渐近正态性 (参见文献 [37, 39]). 此外, 还有一些有趣性质. 例如, 在一定条件下, 使用倾向得分估计值进行匹配得到的平均因果作用估计的方差比使用倾向得分的真实值还小 (参见文献 [37, 39]). 文献 [40–42] 详细回顾了匹配方法在应用中的其他话题, 例如, 方差估计的方法、有放回 (replacement) 匹配与无放回匹配的比较、精确匹配、最近邻法匹配 (nearest neighbor matching)、检验匹配数据平衡性的方法、统计软件实现等.

3.4 逆概加权估计和回归估计

除了匹配, 倾向得分还经常用在逆概加权估计 (inverse probability weighted estimation) 中. 给定可忽略性假定, 容易证明,

$$\mathrm{E}(Y_x)=\mathrm{E}\left\{\frac{\delta(X=x)Y}{\mathrm{pr}(X=x\mid V)}\right\},$$

其中 $\delta(\cdot)$ 是示性函数. 据此, 可以通过拟合一个倾向得分模型 $\pi(V;\alpha)=\mathrm{pr}(X=1\mid V;\alpha)$ 来估计平均因果作用. 倾向得分模型满足相应的矩方程 $\mathrm{E}\{X-\pi(X;\alpha)\mid V\}=0$, 因此可以用经典的方法, 如广义矩估计 (generalized method of moments, GMM [43, 44]) 来估计未知参数 α. 得到参数估计 $\widehat{\alpha}$ 后, 平均因果作用的逆概加权估计为

$$\widehat{\phi}_{\rm ipw}=\sum_{i=1}^{n}\frac{X_iY_i}{\pi(V_i;\widehat{\alpha})}-\sum_{i=1}^{n}\frac{(1-X_i)Y_i}{1-\pi(V_i;\widehat{\alpha})}.$$

回归估计 (regression-based estimator) 需要建立一个对结果变量的回归模型, $\mathrm{E}(Y\mid X,V)=m(X,V;\gamma)$. 为了估计平均因果作用, 需要先估计该模型的参数. 注意到该回归模型满足矩方程 $\mathrm{E}\{Y-m(X,V;\gamma)\mid X,V\}=0$, 可以用经典的估计矩方程的方法来估计未知参数 γ. 在得到参数估计 $\widehat{\gamma}$ 后, 平均因果作用的回归估计为

$$\widehat{\phi}_{\rm reg}=\frac{1}{n}\sum_{i=1}^{n}X_im(1,V_i;\widehat{\gamma})-\frac{1}{n}\sum_{i=1}^{n}(1-X_i)m(0,V_i;\widehat{\gamma}).$$

给定可忽略性假定和一定的正则条件, 容易证明, 如果回归模型和倾向得分模型分别正确, 并且 $\widehat{\gamma}$ 和 $\widehat{\alpha}$ 分别是 γ 和 α 的相合估计, 那么 $\widehat{\phi}_{\rm reg}$ 和 $\widehat{\phi}_{\rm ipw}$ 是平均因果作用的相合估计 (参见文献 [45,46]). 回归估计和逆概加权估计方法被广泛应用在流行病学、生物医学、社会学和经济学的因果推断中. 这两种方法简捷明了, 但是对模型特别敏感. 当所需要的回归模型或者倾向得分模型不正确时, 往往导致较大偏差. 下面介绍的双稳健估计则利用回归估计与逆概加权估计互补, 有效地缓解这两种方法各自的缺点.

3.5 双稳健估计

文献 [46–48] 在研究缺失数据时提出了双稳健估计方法 (doubly robust estimation). 这种方法把回归估计和逆概加权估计结合起来, 并具有双稳健性质: 只要回归模型和倾向得分模型中的一个模型正确, 那么双稳健估计就有相合性.

双稳健估计同时需要一个回归模型 $m(X,V;\gamma)=\mathrm{E}(Y\mid X,V;\gamma)$ 和一个倾向得分模型 $\pi(V;\alpha)=\mathrm{pr}(X=1\mid V;\alpha)$. 估计未知参数 (α,γ) 的方法如第 3.4 小节所述. 得到参数估计 $(\widehat{\alpha},\widehat{\gamma})$ 后, 平均因果作用的双稳健估计为

$$\begin{aligned}\widehat{\phi}_{\rm dr}=&\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\pi(V_i;\widehat{\alpha})}Y_i-\frac{1}{n}\sum_{i=1}^{n}\frac{1-X_i}{1-\pi(V_i;\widehat{\alpha})}Y_i\\&+\frac{1}{n}\sum_{i=1}^{n}\left\{1-\frac{X_i}{\pi(V_i;\widehat{\alpha})}\right\}m(1,V_i;\widehat{\gamma})-\frac{1}{n}\sum_{i=1}^{n}\left\{1-\frac{1-X_i}{1-\pi(V_i;\widehat{\alpha})}\right\}m(0,V_i;\widehat{\gamma}).\end{aligned}\tag{3.1}$$

上式的第一行等于逆概加权估计, 第二行是对逆概加权估计的一个纠偏项, 由逆概的残差和回归估计构成. 如果倾向得分模型正确, 那么逆概加权估计有相合性, 并且当样本量增加时第二行中的纠偏项趋于零. 这是因为, 根据大数定律, 上式中的第三项收敛到

$$\mathrm{E}\left[\left\{1-\frac{X}{\pi(V;\alpha)}\right\}m(1,V;\gamma^*)\,\middle|\,V\right]=\mathrm{E}\left[m(1,V;\gamma^*)\mathrm{E}\left\{\frac{\pi(V;\alpha)-X}{\pi(V;\alpha)}\,\middle|\,V\right\}\right]=0,$$

其中 $\gamma^*=\mathrm{plim}\,\widehat{\gamma}$, 表示当样本量趋于无穷时 $\widehat{\gamma}$ 依概率收敛的极限值. 同理第四项也收敛到 0. 因此, $\widehat{\phi}_{\rm dr}$ 在倾向得分模型 $\pi(V;\alpha)$ 正确时有相合性. 注意, 当回归模型 $m(X,V;\gamma)$ 错误时, 上面的推导仍然成立.

(3.1) 中的双稳健估计还可以等价地表示为

$$\begin{aligned}\widehat{\phi}_{\rm dr}=&\frac{1}{n}\sum_{i=1}^{n}m(1,V_i;\widehat{\gamma})-\frac{1}{n}\sum_{i=1}^{n}m(0,V_i;\widehat{\gamma})\\&+\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\pi(V_i;\widehat{\alpha})}\{Y_i-m(1,V_i;\widehat{\gamma})\}-\frac{1}{n}\sum_{i=1}^{n}\frac{1-X_i}{1-\pi(V_i;\widehat{\alpha})}\{Y_i-m(0,V_i;\widehat{\gamma})\}.\end{aligned}\tag{3.2}$$

(3.2) 中的第一行是回归估计, 第二行是对回归估计的一个纠偏项. 如果回归模型正确, 那么回归估计有相合性, 而且可以证明, 无论倾向得分模型正确与否, 当样本量增加时 (3.2) 中第二行中的纠偏项趋于零, 因此, $\widehat{\phi}_{\rm dr}$ 在回归模型正确时有相合性, 而不需要倾向得分模型正确.

综上, $\widehat{\phi}_{\rm dr}$ 具有双稳健性质. 相比于回归估计和倾向得分估计, 双稳健估计提供了更多减少估计偏差的机会. 由于双稳健估计能有效地减小模型错误导致的偏差, 这种方法越来越广泛应用在缺失数据分析和因果推断中. 关于双稳健估计理论性质的进一步研究和总结参见文献 [49–51]. 在假定可忽略

性的条件下, 双稳健估计的构造方法被迅速扩展 (参见文献 [52–55]). 但是, 在可忽略性假定不满足时, 倾向得分模型依赖未完全观测的潜在结果, 联合分布的识别性得不到保证. 此时, 双稳健估计的构造比较困难, 通常需要很强的倾向得分模型假定 (参见文献 [56]). 最近, 文献 [57–59] 提出使用辅助变量构造双稳健估计的方法. 这些方法使用辅助变量提高联合分布的识别性和估计倾向得分模型, 从而避免了过强的模型假定.

但是要注意到, 当两个模型都不正确时, (3.1) 或 (3.2) 的双稳健估计可能会比回归估计和逆概加权估计的偏差更大. 当出现特别大或者特别小的倾向得分时, 偏差会被放大, 甚至出现不合理的估计结果. 例如, 对一个取 0 和 1 值的结果变量, 当两个模型都错误时, 双稳健估计可能会得到大于 1 的结果, 详细的讨论和例子参见文献 [60]. 针对这些缺陷的改进方法, 参见文献 [61–63].

4 未观测的混杂因素和潜在可忽略性

匹配、逆概加权、回归和双稳健估计方法的重要前提是可忽略性假定. 然而, 在实际研究中, 如果有重要背景变量未被观测、测量误差或者选择偏差, 就有潜在的未观测的混杂因素, 可忽略性假定可能不成立, 前一节介绍的统计推断方法在出现未观测的混杂因素时就有偏差.

当存在未被观测的混杂因素时, 更合理的假定是潜在可忽略性: 存在未被观测的变量 U 满足 $Y_x \perp\!\!\!\perp X \mid (U,V)$, 其中 V 为观测的混杂因素. 为了记号简略, 本节省略掉观测的混杂因素 V.

假定 4.1 (潜在可忽略性) $Y_x \perp\!\!\!\perp X \mid U$.

在 U 是常数时, 此假定退化为可忽略性假定. 在潜在可忽略性假定下,

$$\mathrm{E}(Y_x) = \mathrm{E}\{\mathrm{E}(Y \mid X = x, U)\} \neq \mathrm{E}(Y \mid X = x).$$

如果 U 没有被观测, 那么 $\mathrm{E}(Y \mid X = x, U)$ 一般不能由观测数据识别, 因此, $\mathrm{E}(Y_x)$ 的识别性不能保证. 如果用 $\mathrm{E}(Y \mid X = x)$ 来估计 $\mathrm{E}(Y_x)$ 就产生偏差. 在潜在可忽略性假定下, 辅助变量经常被用来帮助识别因果作用和消除混杂偏倚. 辅助变量通常只与 (X, Y, U) 三个变量的一个子集相关, 因此引入一些条件独立性帮助识别因果作用. 本节介绍在潜在可忽略性假定下用来消除混杂偏差的两种方法, 一种是常用的工具变量 (instrumental variable) 方法, 一种是最近引起人们注意的阴性对照变量 (negative control variable) 方法.

4.1 工具变量估计

用 Z 表示一个工具变量, U 表示未观测的混杂因素, 工具变量的定义包含三个核心条件:

假定 4.2 (i) $Z \perp\!\!\!\perp Y \mid (X, U)$; (ii) $Z \perp\!\!\!\perp U$; (iii) $Z \not\perp\!\!\!\perp X$.

当存在完全观测的协变量 V 时, 以上的三个条件改为在给定 V 条件下成立. 条件 (i) 表示工具变量 Z 对 Y 无直接作用; 条件 (ii) 表示工具变量与未观测的混杂因素独立; 条件 (iii) 表示工具变量与处理相关. 当工具变量 Z 是发生在 X 之前的一种处理或暴露变量时, 则可以定义潜在接受的处理 X_z 和潜在结果 $Y_{z,x}$. 例如, 在临床试验中, 由于患者存在不依从性, 医生随机分配的处理 (assignment) Z 会影响, 但可能不同于患者实际接受的处理 (treatment) X; 患者治疗结果变量 Y 只依赖于实际接受的处理, 而不受处理分配的影响. 此时, 处理分配 Z 可以作为工具变量, 满足下面的等价于假定 4.2 的条件:

假定 4.3 (i) $Y_{z,x} = Y_x$; (ii) $Y_{z,x} \perp\!\!\!\perp Z$; (iii) $Z \not\perp\!\!\!\perp X$.

给定这三个条件, 平均因果作用仍然不可识别, 但是可以求平均因果作用的上下界, 也称为部分可识别性 (partial identification). 文献 [64–66] 研究了平均因果作用在条件 (i)–(iii) 下的上下界. 但是这些界通常仍然比较宽, 在实际研究中不足以得到很确定的因果推论. 因此, 人们通常引入一些模型假定, 这些假定能进一步缩短因果作用的界或者识别因果作用.

除了需要条件 (i)–(iii), 有两种常用的假定被用在工具变量的分析中 (参见文献 [67, 第 16 节]): 一种是因果作用的同质性假定 (effect homogeneity), 一种是单调性假定 (monotonicity). 同质性假定最常用的版本即为在经济学和社会学中广泛应用的结构方程模型 (structural equation model). 工具变量被用来估计结构方程中处理或暴露变量的回归系数 (参见文献 [68–70]). 例如, 连续型结果变量的线性模型为

$$Y = \beta_0 + \beta_1 X + U, \tag{4.1}$$

其中 U 是未观测的混杂因素, β_1 表示在其他因素 (U) 不变的情形下, X 每增加一个单位对 Y 的作用 (ceteris paribus effect). 这个方程实际上假定 X 对 Y 的作用在所有人当中是一个常数. 这个方程也可以等价地用潜在结果表示为 $Y_x = \beta_0 + \beta_1 x + U$. 此模型隐含了潜在可忽略性假定 4.1. 在此模型下, β_1 代表平均因果作用 $\mathrm{E}(Y_1 - Y_0)$. 由于存在未观测的混杂因素, 因此仅从结构方程不能识别 β_1. 例如, 当 $\mathrm{E}(U \mid X) \neq 0$ 时, β_1 的最小二乘估计有偏. 但是利用工具变量, 可以识别 β_1. 利用一个满足假定 4.2 的工具变量 Z, 可以验证 $\beta_1 = \sigma_{yz}/\sigma_{xz}$. 给定观测数据, 把样本协方差 $\widehat{\sigma}_{yz}$ 和 $\widehat{\sigma}_{xz}$ 代入, 即得到 β_1 的工具变量估计 (instrumental variable estimator)

$$\beta_1^{\mathrm{iv}} = \frac{\widehat{\sigma}_{yz}}{\widehat{\sigma}_{xz}}. \tag{4.2}$$

即使 $\mathrm{E}(U \mid X) \neq 0$, 在假定 (i)–(iii) 和一定的正则条件下, 可以证明 β_1^{iv} 的相合性和渐近正态性 (参见文献 [70, 第 15 节]). 工具变量方法有效地缓解了未观测的混杂因素导致的偏差.

比常数作用假定稍弱的是文献 [71] 提出的无修正作用假定: $\mathrm{E}(Y_1 - Y_0 \mid Z, X) = \mathrm{E}(Y_1 - Y_0 \mid X)$. 在此假定下, 可以识别处理组的平均因果作用 $\mathrm{E}(Y_1 - Y_0 \mid X = 1)$. 同质性假定在实际中难以论证是否成立, 此外, 结构方程模型需要参数模型, 对模型的错误设定敏感. 文献 [71] 讨论了这些模型和假定在实际应用中, 特别是在流行病学研究中的局限性.

不同于同质性假定, 单调性假定要求 Z 对 X 的作用单调. 单调性假定在很多情形下更合理. 例如, 在研究药效的临床试验中, 由于患者存在不依从性, 患者实际接受的处理 (treatment, X) 和医生分配的处理 (assignment, Z) 不完全相同, 但可以合理地假定单调性: 所有个体潜在接受的处理都满足 $X_{z=1} \geqslant X_{z=0}$. 在此假定下, 文献 [72, 73] 证明了可以使用 Z 作为工具变量识别依从组 (即满足 $X_{z=1} = 1$ 和 $X_{z=0} = 0$ 的个体) 的平均因果作用 (compliers average causal effect): $\mathrm{E}(Y_{x=1} - Y_{x=0} \mid X_{z=1} = 1, X_{z=0} = 0)$.

工具变量已被广泛应用在经济学、社会学、流行病学和生物统计学的观察性研究中, 这方面的研究数不胜数. 与因果推断密切相关的问题包括, 半参数和非参数结构方程模型的估计 [74, 75]、在二值结果变量模型中的应用 [76, 77]、Mendel 随机化研究 (Mendelian randomization) 中的应用 [78] 和弱工具变量的问题 [79, 80]. 工具变量最近的综述性文章参见文献 [76, 81].

4.2 阴性对照变量方法

工具变量估计对假定 4.2(i)–4.2(iii) 很敏感. 假定 4.2 的条件 (i) 需要有专业知识保证工具变量对结果变量没有直接作用, 条件 (iii) 可以用观测数据检验, 但是条件 (ii) 难以验证, 因为混杂因素 U 没

有观测到. 用 plim β_1^{iv} 表示当样本量趋于无穷时 β_1^{iv} 依概率收敛的极限. 当条件 (ii) 不满足时, 工具变量估计的渐近偏差是

$$\text{plim}\,\beta_1^{\text{iv}} - \beta_1 = \frac{\sigma_{uz}}{\sigma_{xz}}. \tag{4.3}$$

由此可见, 当 $\sigma_{uz} \neq 0$ 时, 工具变量估计不相合, 而且偏差会由于 σ_{xz} 过小而被放大很多倍. 我们介绍使用阴性对照变量 (negative control variable) 方法解决这些问题.

阴性对照变量是与混杂因素 U 相关, 但与处理 X 或结果变量 Y 无因果关系的辅助变量. 阴性对照变量分为两种: 阴性对照暴露和阴性对照结果. 前者是一个辅助的暴露变量, 但是对关心的结果没有直接的因果作用; 后者是一个辅助的结果变量, 但是不受暴露变量的影响. 这些特点可以严格地表述如下.

假定 4.4 (阴性对照结果, negative control outcome) 一个结果变量 W 称为一个阴性对照结果, 如果它满足 $W \perp\!\!\!\perp X \mid U$ 和 $W \not\perp\!\!\!\perp U$.

假定 4.5 (阴性对照暴露, negative control exposure) 一个暴露变量 Z 称为一个阴性对照暴露, 如果它满足 $Z \perp\!\!\!\perp Y \mid (U, X)$ 和 $Z \perp\!\!\!\perp W \mid (U, X)$.

除了要求阴性对照变量 Z 和 W 与处理或结果变量无直接的因果关系, 上面的定义还要求 Z 与 (W, Y) 之间的混杂因素和 X 与 (W, Y) 之间的混杂因素相同. 当存在完全观测的协变量 V 时, 上述定义中的条件独立性需要给定 V. 阴性对照暴露的定义类似工具变量中的无直接作用条件, 但是对 (Z, U) 的相关性不做要求, 因此, 工具变量可看作阴性对照暴露的特例.

在流行病学中, 阴性对照变量在很长时间内被用作检测混杂因素是否存在 (参见文献 [82–87]). 例如, 在一项关于突发压力对心脏病的危害的研究中, Trichopoulos 等 [88] 发现, 在雅典 1981 地震后短时间内由心脏病导致的死亡增加, 为了验证是否存在未观测的混杂因素, 他们又分析了癌症死亡数据, 发现癌症导致的死亡没有明显增加. 这样的结果表明, 没有明显的混杂作用, 也间接地支持了地震带来的突发压力增加心脏病风险的结论. 在这一例子中, 癌症死亡率作为阴性对照结果用来检验是否存在未观测的混杂. 文献 [86, 89, 90] 回顾了使用阴性对照变量检测混杂的方法. 在流行病学和生物统计研究中, 阴性对照变量还被用来校正混杂导致的偏差 [87, 91–93], 但这些方法通常需要很强的模型假定. 如何使用阴性对照变量得到确定性的因果推论和识别因果作用, 目前的研究很少.

最近, 文献 [94–96] 系统地研究了用阴性对照变量识别因果作用的方法和所需要的条件. 文献 [96] 引入的混杂桥函数 (confounding bridge function) 为使用阴性对照变量识别因果作用奠定了基础.

假定 4.6 存在一个函数 h, 对所有 x, 都有 $\text{E}(Y \mid U, X = x) = \text{E}\{h(W, X = x) \mid U\}$.

当存在协变量 V 时, 该假定为 $\text{E}(Y \mid U, V, X = x) = \text{E}\{h(W, V, X = x) \mid U, V\}$. 混杂桥函数描述的是混杂因素对关心的结果变量与对阴性对照结果的作用之间的关系. 例如, 当 $\text{E}(Y \mid U, X) = \beta_1 X + U$ 且 $\text{E}(W \mid U)$ 是 U 的线性模型时, 混杂桥函数是一个线性可加函数 $h(W, X; \gamma) = \gamma_0 + \gamma_1 X + \gamma_2 W$, 其中的参数由数据产生机制 $\text{E}(Y \mid U, X)$ 和 $\text{E}(W \mid U)$ 决定.

在假定 4.1 和 4.4–4.6 下, 容易得到

$$\text{E}(Y_x) = \text{E}\{h(W, X = x)\}. \tag{4.4}$$

可见, 使用混杂桥函数, 潜在结果的均值可以由阴性对照结果的分布表示出来. 如果已知 h 函数, 那么 $\text{E}(Y_x)$ 可以根据 (4.4) 识别, 进而平均因果作用可以识别. 例如, 当 h 满足线性模型时, $h(W, X; \gamma) = \gamma_0 + \gamma_1 X + \gamma_2 W$, 那么 γ_1 就等于平均因果作用. 在实际研究中, 如果 h 是未知函数, 可以利用一个阴性对照暴露识别 h.

由假定 4.5 和 4.6 可得

$$\mathrm{E}(Y \mid Z, X) = \mathrm{E}\{h(W, X) \mid Z, X\}. \tag{4.5}$$

(4.5) 中只涉及观测数据的分布和未知的混杂桥函数 h, 因此提供了识别 h 的基础. 此外, 还需要如下的完备性条件.

假定 4.7 (完备性) 对任意 x, 条件分布 $\mathrm{pr}(W \mid Z, X = x)$ 满足: 对所有的平方可积函数 g, 如果 $\mathrm{E}\{g(W) \mid Z, X = x\} = 0$ 几乎处处成立, 那么 $g(W)$ 几乎处处为零.

完备性条件是研究识别性问题中普遍需要的条件. 作为特例, 对一个二值的混杂因素, 只要 $U \not\perp\!\!\!\perp Z \mid X = x$ 对任意 x 成立, 那么假定 4.7 一定成立. 此外, 完备性条件在很多模型下成立, 如指数族分布 [74]. 关于完备性条件的讨论参见文献 [74, 97–100].

定理 4.1 文献 [96] 在假定 4.1 和 4.4–4.7 下, 方程 (4.5) 有唯一解 h; 把此解代入 (4.4), 则 $\mathrm{E}(Y_x)$ 可识别.

至此, 通过引入混杂桥函数, 使用阴性对照变量已经可以识别 $\mathrm{E}(Y_x)$ 和平均因果作用.

在使用实际数据估计因果作用时, 可以对混杂桥函数建立参数化模型 $h(W, X; \gamma)$, 然后用广义矩方法或者两步最小二乘估计 γ 和平均因果作用. 以线性模型为例, 假设 $\mathrm{E}(Y \mid X, U) = \beta_1 X + U$, 那么, 我们采用线性可加的混杂桥函数 $h(W, X; \gamma) = \gamma_0 + \gamma_1 X + \gamma_2 W$, 并用两步最小二乘方法估计 β_1: 第 1 步用 W 对 (Z, X) 做回归得到 $\widehat{W}$, 第 2 步用 Y 对 $(\widehat{W}, X)$ 做回归. 那么第 2 步回归中得到的 X 的系数估计即为 β_1 的相合估计, 也是平均因果作用的相合估计. 还可以证明, β_1 的两步最小二乘方法估计等价于

$$\beta_1^{\mathrm{nc}} = \frac{\widehat{\sigma}_{xw}\widehat{\sigma}_{zy} - \widehat{\sigma}_{xy}\widehat{\sigma}_{zw}}{\widehat{\sigma}_{xw}\widehat{\sigma}_{xz} - \widehat{\sigma}_{xx}\widehat{\sigma}_{zw}}.$$

此式可看作 (4.2) 的推广. 当 $Z \perp\!\!\!\perp U$, 即 Z 是一个工具变量时, $\mathrm{plim}\widehat{\sigma}_{zw} = 0$, 因此, $\mathrm{plim}\beta_1^{\mathrm{nc}} = \mathrm{plim}\beta_1^{\mathrm{iv}} = \beta_1$; 当 $Z \not\perp\!\!\!\perp U$ 时, $\mathrm{plim}\beta_1^{\mathrm{iv}} \neq \beta_1$, 但是, 如果 $\mathrm{E}(W \mid U)$ 是 U 的线性模型, 那么仍然有 $\mathrm{plim}\beta_1^{\mathrm{nc}} = \beta_1$. 结果, 使用阴性对照 W 有效地缓解了工具变量估计在假定 $Z \perp\!\!\!\perp U$ 不成立时的偏差. 除了参数化模型, 在复杂问题中还可以对 $h(W, X)$ 建立半参数化或者非参数化的模型进行估计, 这些都属于一大类条件矩模型 (conditional moment restriction model) 的估计方法, 已有很多标准的做法, 参见文献 [74, 75].

5 替代指标与中介分析

在科学研究中, 特别是在医学和生物学试验中, 当感兴趣的终点指标 (endpoint) 难以观测时, 常会取而代之观测替代指标 (surrogate) 或标记物 (marker). 如何寻找或确定替代指标是一个尚未解决的问题. 在医学临床试验中一些常用的替代指标遭到了质疑, 文献 [101–104] 指出了在临床试验中由于使用替代指标错误评价治疗效果的实例, 例如, AIDS (acquired immune deficiency syndrome) 病临床试验中采用的替代指标 CD4、预防骨质疏松采用的替代指标骨密度等.

目前有若干种确定替代指标的准则. 最直观的准则是要求替代指标与终点指标有强相关性. 但是, 强相关的替代指标不意味着因果关系. 例如, 小孩鞋子尺寸与记忆的单词量有很强的正相关性, 增加鞋子尺寸并不能增加单词量. 文献 [105] 提出了统计替代指标的准则, 除了要求替代指标与终点指标相关之外, 还要求给定替代指标下处理 (treatment) 与终点指标条件独立. 统计替代指标只是切断了处理与终点指标的相关关系, 不能切断因果关系. 文献 [106] 提出了切断处理与终点指标之间因果关系的主替代指标 (principal surrogate), 处理对替代指标没有因果作用的话, 处理对终点指标就没有因果作

用, 即满足因果必要性. 文献 [3] 利用因果网络图提出了强替代指标的准则, 要求强替代指标切断处理到终点指标的因果路径. 文献 [107] 提出了替代指标悖论, 即处理 (或称治疗) 对替代指标有正的因果作用, 并且替代指标对终点指标也有正的因果作用, 但是该处理对终点指标有负的因果作用. 例如, 任何一个人假若心律正常一定比心律不正常活得更长, 某种药可以显著纠正心律失常, 但是这种药反而减少患者的寿命 [108]. 文献 [107] 指出了前面所述的准则都不能避免替代指标悖论的发生. 令人惊讶的是, 主替代指标和强替代指标有非常严格的要求条件, 但是它们仍然不能避免替代指标悖论; 替代指标悖论从根基上完全动摇了替代指标和标记物的价值 [109]. 在文献 [109] 之后, Elliott 研究组、Joffe 和 Pearl 讨论了替代指标悖论的重要性, 共同认识到非常有必要重新定义和评价合理的替代指标及其准则, 以避免替代指标悖论现象的发生.

5.1 替代指标悖论

文献 [108] 讲述了有关治疗心律失常药物导致美国数万人死亡的重大药物灾难. 该药物灾难描述了一个替代指标悖论的真实例子. 关于心律失常疾病, 医生的知识是心律失常是猝死的危险因素, 纠正心律失常能够预防猝死. 因此, "纠正心律失常" 作为一个替代指标, 美国 FDA (Food and Drug Administration) 批准了几种药物 (Enkaid、Tambocor 和 Ethmozine). 后来通过上市后追踪研究发现, 尽管这些药物能有效地纠正心律失常, 但是反而增加了死亡率.

为了理解为什么会发生替代指标悖论, 下面给出一个数值例子说明主替代指标和强替代指标可能会发生替代指标悖论的现象. 令 X 为二值的处理, 1 表示接受一种新的处理, 0 表示接受对照处理. S 表示心律失常是否纠正, 1 表示纠正了, 0 表示未纠正. S_x 表示接受处理 $X=x$ 情形下是否纠正心律失常的潜在结果. Y_{sx} 表示在处理 $X=x$ 且心律失常纠正与否 $S=s$ 情形下的潜在生存时间. 假定处理 X 对生存时间 Y 的作用完全通过中间变量 S 起作用, 即 $Y_{sx}=Y_{sx'}=Y_s$. 因此, 当处理 X 对心律 S 没有个体因果作用 $S_{X=1}=S_{X=0}=s$ 时, 处理 X 对生存时间 Y 也没有因果作用 $Y_{s1}=Y_{s0}$, 即 S 是一个主替代指标, 也是一个强替代指标. 进一步假定纠正心律失常能延长每一位患者 i 的生存时间 $Y_{S=0}(i)<Y_{S=1}(i)$. 在这两个假定下, 直观上似乎纠正心律失常可以作为生存时间的一个 "理想" 的替代指标. 从统计意义上说, 如果处理 X 能纠正心律失常 S, 就应该能 "延长" 患者的寿命 Y. 但是, 这个直观是错误的. 我们可以设想一个 100 位心律失常患者的总体 (如表 1 所示) 来说明替代指标悖论的现象. 100 位患者分为四个主分层: $(S_0=0, S_1=0)$、$(S_0=0, S_1=1)$、$(S_0=1, S_1=0)$ 和 $(S_0=1, S_1=1)$, 如第 3 和 4 列所示. 在每一主分层的患者人数如第 2 列所示. 为简单起见, 假定在同一主分层的所有患者有相同的潜在生存时间, 第 5 和 6 列给出了所有主分层中患者的潜在生存时间 (年), 其中有些是先验反事实的. 例如, 在第 1 层的是先验反事实的, 这是因为不管接受对照治疗 $T=0$, 还是接受处理 $X=1$ 都不会得到 $S=1$, 除非采用干预才可能得到. 由第 3 到 6 列可以得到潜在生存时间. 由表 1 的 100 位患者的总体, 可以得到处理 X 对纠正心律失常 S 的平均因果作用为: 假若 100 人都

表 1　100 位心律失常患者的总体

主分层	人数	$S_{X=0}$	$S_{X=1}$	$Y_{S=0}$	$Y_{S=1}$	$Y_{X=0}$	$Y_{X=1}$
1	20	0	0	3	5	3	3
2	40	0	1	6	7	6	7
3	20	1	0	5	8	8	5
4	20	1	1	9	10	10	10

接受处理 $X=1$ 时纠正心律失常的比率与假若 100 人都接受对照 $X=0$ 时纠正心律失常的比率之差

$$\mathrm{ACE}_{X\to S}=\frac{40+20}{100}-\frac{20+20}{100}=\frac{20}{100}>0,$$

表明处理 X 对纠正心律失常 S 有正的因果作用. 而处理 X 对生存时间 Y 的因果作用为: 假若 100 人都接受处理 $X=1$ 时的平均寿命与假若 100 人都接受对照 $X=0$ 时的平均寿命之差

$$\mathrm{ACE}_{X\to Y}=\frac{3\cdot 20+7\cdot 40+5\cdot 20+10\cdot 20}{100}-\frac{3\cdot 20+6\cdot 40+8\cdot 20+10\cdot 20}{100}=-\frac{20}{100}<0,$$

表明处理 X 对患者寿命 Y 有负的因果作用. 这个例子说明了纠正心律失常 S 作为一个主替代指标和强替代指标, 导致了替代指标悖论的现象, 不能正确地评价处理 X 对寿命 Y 的作用. 这个例子说明了因果作用的统计结论不具有传递性. 处理 X 对纠正心律失常 S 的平均因果作用为正, 纠正心律失常 S 对寿命 Y 的个体因果作用为正, 但是处理 X 对寿命 Y 的平均因果作用为负.

上面的例子中存在一个主分层 ($S_{X=0}=1$, $S_{X=1}=0$), 意味着对于该层中的患者, 处理 $X=1$ 反而导致心律失常. 实际上, 即使该主分层不存在, 仍可能会发生替代指标悖论的现象.

替代指标悖论又称为工具变量悖论、中间指标悖论. 该悖论意味着: 工具变量估计可能出现正负符号的悖论现象, 利用中间变量的统计结论不具有传递性. Pearl [110] 也提出了类似的问题, 发表了题为 "科学知识对决策是否有用" 的论文. 例如, 吸烟会提高肺癌的患病率已被公认为科学知识, 某些禁烟政策也确实能减少吸烟人数, 但是, 这些政策可能反而会提高肺癌患病率.

Yule-Simpson 悖论揭示了从数据得到因果结论的困难, 甚至不可能, 除非采用随机化试验. 替代指标悖论进一步揭示了即使采用随机化试验能得到因果结论, 但是, 统计的因果结论不能用于推理, 统计结论不具有传递性. 例如, 两个随机化试验得到统计因果结论: 变量 X 能提高变量 S, 变量 S 能提高变量 Y, 但是根据这两个因果结论不能推出变量 X 提高变量 Y.

5.2 避免替代指标悖论的准则

我们寻找替代指标准则的目标是, 不必观测终点指标 Y, 而只需观测替代指标 S, 就可以用处理 X 对观测的替代指标 S 的因果作用的正负符号来预测处理 X 对未观测的终点指标 Y 的因果作用的正负符号. 文献 [107, 109, 111, 112] 提出和探讨了一致性替代指标的准则, 其目的是避免替代指标悖论现象. 设 X 是二值的随机化处理, S 是图 1 定义的强替代指标. 该强替代指标阻断了处理 X 到终点指标 Y 的路径. 通常替代指标不能随机化, 因此可能存在替代指标与终点指标之间的混杂因素 U. 处理 X 对终点指标 Y 的分布因果作用 (distributional causal effect, DCE) 定义为

$$\mathrm{DCE}_{X\to(Y>y)}=\mathrm{pr}(Y_{X=1}>y)-\mathrm{pr}(Y_{X=0}>y),$$

其中 y 为某给定值.

文献 [107, 111] 证明了, 如果 (1) 任意给定 $U=u$ 下 S 对 Y 有非负的条件 ACE; (2) 任意给定 $U=u$ 下 X 对 S 的条件 DCE 的符号 (非负, 或非正) 不随 u 改变, 那么, 根据 X 对 S 的非负 (非正) DCE 能预测 X 对 Y 的非负 (非正) ACE, 并且根据 X 对 S 的零 DCE 能预测 X 对 Y 的零 ACE. 条

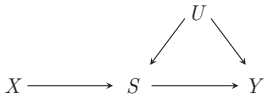

图 1 强替代指标 S 切断处理 X 到终点指标 Y 的因果路径

件 (1) 的意思是, 给定任意 $U=u$ 的条件下替代指标 S 都是终点指标 Y 的一个危险因素 (或保护因素). 例如, 在每个 $U=u$ 的子总体中, 心律失常都是猝死的危险因素. 条件 (2) 的意思是给定任意 u 的条件下处理 X 对替代指标 S 有相同符号的 DCE. 例如, 在任何子总体 $U=u$ 中, 新药对纠正心律失常都没有反作用或都没有正作用. 文献 [107,111] 探讨了很多常用的模型, 包括广义线性模型和比例风险模型, 它们都满足这些条件.

因为不能观测到混杂因素 U, 即使观测到终点指标 Y, 这两个条件也不能用数据进行检验. 另外, 强替代指标要求从处理 X 到终点指标 Y 的所有路径都被中间变量 S 阻断, 这意味着没有从处理 X 到终点指标 Y 的直接作用. 为了消除上述两个缺点, 文献 [112] 提出了基于关联性条件的一致替代指标准则. 这个准则要求未观测终点指标和其他变量之间关联性的先验知识. 如果在以前的研究中观测到终点指标 Y 的话, 这些关联性的先验知识条件是可以检验的, 并且它们不要求没有对终点指标的直接处理作用.

目前, 大部分关于替代指标的讨论都是围绕单个替代指标的. 主代理和强代理都要求单个替代指标能切断处理到终点指标的因果路径. 在实际中, 处理到终点指标可能会有多条因果路径, 需要多个替代指标才能切断所有路径. 例如, 评价某种治疗对 AIDS 病的疗效, 不仅需要看该治疗提高 CD4 的作用, 还应该看该治疗减少 HIV-1 RNA (ribonucleic acid) 的作用. 文献 [113] 提出了多替代指标的准则, 探讨了从处理 X 到终点指标 Y 的因果路径上有多个中间变量的情形, 例如, S_1 和 S_2 是两个可能的替代指标. 因为替代指标 S_1 和 S_2 没有随机化分配, 所以, 可能存在没有观测的混杂或是混杂向量 U 同时影响 S_1、S_2 和终点指标 Y. 两个替代指标的因果网络可以用图 2 描述, S_1 与 S_2 之间的双箭头表示 S_1 指向 S_2 或者 S_2 指向 S_1; S_{1x}、S_{2x} 和 Y_{sx} 表示潜在结果.

令 $\boldsymbol{x}=(x_1,x_2,x_3,\ldots,x_n)$ 和 $\boldsymbol{y}=(y_1,y_2,y_3,\ldots,y_n)$. $\boldsymbol{x}\preceq\boldsymbol{y}$ 表示 $x_i\leqslant y_i$, $i=1,2,3,\ldots,n$. 如果当 $\boldsymbol{x}\preceq\boldsymbol{y}$ 时, 就有 $\boldsymbol{\phi}(\boldsymbol{x})\leqslant(\geqslant)\ \boldsymbol{\phi}(\boldsymbol{y})$, 那么称函数 $\boldsymbol{\phi}(\boldsymbol{x})$ 是增加的 (或是减少的). 一个集合 $U\subseteq\mathbb{R}^n$ 称为是增加的 (下降的), 如果 $\boldsymbol{y}\succeq(\preceq)\ \boldsymbol{x}$ 和 $\boldsymbol{x}\in U$, 就有 $\boldsymbol{y}\in U$. 如果两个随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 满足如下的条件: 对于所有的增加集合 U, 有 $\mathrm{pr}(\boldsymbol{X}\in U)\leqslant\mathrm{pr}(\boldsymbol{Y}\in U)$, 那么, 称 $\boldsymbol{X}$ 在通常随机序 (stochastic turn) 情形下小于 $\boldsymbol{Y}$ (记为 $\boldsymbol{X}\leqslant_{\mathrm{st}}\boldsymbol{Y}$).

下面介绍替代指标向量的条件, 在这些条件下处理 X 对替代指标向量的因果作用的符号可以预测处理 X 对于终点指标 Y 的因果作用的符号. 记 $f(s_1,s_2,x)=\mathrm{E}(Y\mid S_1=s_1,S_2=s_2,X=x)$. 文献 [113] 给出了下面的结果: 如果下面 3 个条件成立: (1) $f(s_1,s_2,1)$ 或者 $f(s_1,s_2,0)$ 是 s_1 和 s_2 上的增函数; (2) 对于所有的 s_1 和 s_2, 有 $f(s_1,s_2,1)\geqslant f(s_1,s_2,0)$; (3) $(S_{11},S_{21})\geqslant_{\mathrm{st}}(S_{10},S_{20})$, 那么, 处理 X 对于终点指标 Y 的平均因果作用是非负的. 如果该对照 $X=0$ 在先前的临床试验中使用过, 条件 (1) 是可以检验的. 当满足条件独立 $Y\perp\!\!\!\perp X\mid(S_1,S_2)$ 时, 条件 (2) 也成立, 意味着多替代指标切断了处理 X 与终点指标 Y 的关联性, 因此, 这个条件是 Prentice 的单一统计替代指标在多替代指标情形下的扩展. 条件 (3) 表示处理 X 对替代指标向量 (S_1,S_2) 有正的因果作用. 在正态和二值替代指标的情形下, 条件 (3) 可以简化为处理 X 分别对单个替代指标 S_i 的因果作用的条件 (参见文献 [113]). 由这个结果可以看出,

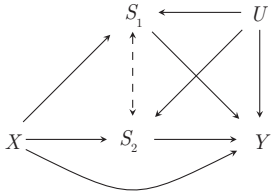

图 2　两个替代指标的因果图

Prentice 的统计替代指标准则需要增加条件 (1) 才能避免替代指标悖论, 而条件 (1) 可以用先前的临床数据进行检验. 与文献 [107, 111] 的条件相比, 假若在核实研究或以前的研究中观测过终点指标 Y, 这些条件可以用数据进行检验. 另外, 这里不要求替代指标向量 (S_1, S_2) 切断从处理 X 到终点指标 Y 的所有路径, 可能还存在从处理 X 到终点指标 Y 的直接作用.

为了避免替代指标悖论的发生, 单个替代指标准则和多替代指标准则都要求一些用数据不可检验的条件, 确定替代指标的准则仍然是一个有待解决的、具有挑战性的重要问题.

6 因果网络及其结构学习

Pearl 教授获得了 2011 年图灵奖, 他的主要贡献是提出了因果网络图模型, 开创了多因素之间的概率因果推理方法, 在人工智能不确定性推理方面取得了突破性进展. 其影响已超出人工智能和计算机科学领域, 乃至影响了人类推理和科学哲学的范畴. 目前, 因果网络方法已经被广泛应用于众多的科学领域. Pearl 基于 Bayes 网络提出了外部干预的概念, 用外部干预的概念对因果关系给出了一种形式化方法, 建立了因果网络模型 [10, 11]. 因果网络描述多个变量之间相互的因果关系, 网络图中每个节点表示一个变量, 节点之间的有向箭头表示由原因到结果的因果关系, 或者变量之间的数据生成过程. 突破了传统统计推断从数据发现相关关系的禁锢, 开创了从数据中发现因果关系及数据产生机制的方法论, 为探索从数据中发现蕴藏在数据中的 “为什么” 建立了基础. 文献 [114] 提出了概率图模型的统计推断和计算方法及其在专家系统中的应用, 建立了大规模因果网络和 Bayes 推断的基础, 取得了专家系统和人工智能中不确定性概率推理的突破性进展.

尽管控制随机试验是发现因果关系的首选方法, 但在实际场景中, 控制试验由于道德、成本和技术等多方面因素, 往往是不可行的. 为此, 众多学者致力于从纯观测数据中发现因果关系. 研究表明, 在特定假设下, 随机变量间的部分或完整因果关系可以从观测数据中还原 [11].

6.1 因果网络

文献 [10, 11, 115, 116] 详细描述了因果网络图, 探讨由观察性研究得到的数据进行因果推断的统计方法. 一个图 $G = (V, E)$ 由节点集合 $V = \{X_1, X_2, \ldots, X_n\}$ 和一个边集合 E 组成. 两个节点之间的一条无向边记为 (X_i, X_j), 一条由 X_i 指向 X_j 的有向边记为 $\langle X_i, X_j\rangle$. 如果所有的边都是无向边, 称该图是一个无向图. 如果所有的边都是有向边, 称该图是一个有向图. 一条从节点 X_i 到另一节点 X_j 的路径 p 是由从 X_i 开始到 X_j 为止、依次有边相连、中间无重复节点的节点和边组成, 而不管边的方向. 如果该路径上所有边的方向都是朝向 X_j, 则称该路径是从 X_i 到 X_j 的有向路径. 一条从 X_i 到 X_i 的有向路径称为一个有向环. 一个没有环的有向图称为有向无环图 (directed acyclic graph, DAG).

令每个节点表示一个随机变量. 令 pa_i 表示变量 X_i 的父节点变量的集合. 每个节点的取值由它的父节点的函数确定 $X_i = f_i(pa_i, \varepsilon_i)$, 其中 ε_i 为不影响网络内部其他节点 $\{X_j, \forall\, j \neq i\}$ 的残余变量. 一般地, 给定一个有向无环图, 随机向量 $(X_1, \ldots, X_n)$ 的联合概率分布为 $\mathrm{pr}(x_1, \ldots, x_n) = \prod_i^n \mathrm{pr}(x_i \mid pa_i)$, 其中 $\mathrm{pr}(\cdot \mid \cdot)$ 表示条件概率. 图 3 给出了一个因果网络的例子, X_4 的父节点集合为 $\{X_2, X_3\}$, 每个变量由它的父节点的函数确定:

$$X_1 = f_1(\varepsilon_1), \quad X_2 = f_2(X_1, \varepsilon_2), \quad X_3 = f_3(X_1, \varepsilon_3), \quad X_4 = f_4(X_2, X_3, \varepsilon_4), \quad X_5 = f_5(X_4, \varepsilon_5).$$

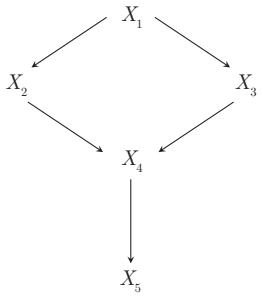

图 3　有向无环图 (DAG)

随机向量 $(X_1,\ldots,X_5)$ 的联合概率分布为

$$\mathrm{pr}(x_1,x_2,x_3,x_4,x_5)=\mathrm{pr}(x_1)\mathrm{pr}(x_2\mid x_1)\mathrm{pr}(x_3\mid x_1)\mathrm{pr}(x_4\mid x_2,x_3)\mathrm{pr}(x_5\mid x_4).$$

对某个变量的外部干预 $X_j=x_j'$ 表示将 $X_j=f_j(pa_j,\varepsilon_j)$ 换成 $X_j=x_j'$, 意味着 X_j 不再受其父节点 pa_j 和 ε_j 的影响, 而强制设定其值为 x_j'. 该干预后的联合分布变为

$$\mathrm{pr}_{x_j'}(x_1,\ldots,x_n)=\delta(x_j=x_j')\prod_{i\neq j}\mathrm{pr}(x_i\mid pa_i),$$

其中 $\delta(\cdot)$ 为示性函数. 特别需要注意的是, $\mathrm{pr}_{x_j'}(x_i)$ 表示在外部干预 $X_j=x_j'$ 下 X_i 的分布, 它不同于条件分布 $\mathrm{pr}(x_i\mid x_j')$. 干预后分布 $\mathrm{pr}_{x_j'}(x_i=1)$ 表示强制所有的人 (包括吸烟的人) 都不吸烟 ($X_j=x_j'$ 表示不吸烟) 的干预下患肺癌 $X_i=1$ 的概率, 而条件分布 $\mathrm{pr}(x_i=1\mid x_j')$ 表示不吸烟人群中 ($X_j=x_j'$) 患肺癌 $X_i=1$ 的概率. 当存在未观测的混杂因素 (X_k, 其影响 X_i 和 X_j) 时, 干预后分布 $\mathrm{pr}_{x_j'}(x_i)$ 是不可识别的. 这是因为强制吸烟的人不吸烟时, 其患肺癌的概率是不可观测的, 可能不同于不吸烟人群的患病概率 $\mathrm{pr}(x_i\mid x_j')$.

在因果网络的框架下, 研究两类问题: 其一是因果作用的可识别性; 另一个是因果网络的学习. 识别因果作用的目的与潜在结果的因果模型的目的是一致的, 即探讨判断混杂因素的准则和研究因果作用的可识别性 [117].

关于因果作用的问题, 利用因果网络可以得到比潜在结果模型更精准的判断混杂因素的准则 [118]. 例如, 一个变量与处理变量和结果变量相关时, 基于潜在结果模型不能判断该变量不是混杂因素; 但是利用因果网络, 如果它不是处理变量和结果变量的共同原因, 那么可以判断它不是混杂因素, 如图 4 中的 Z 和 F. 文献 [117, 118] 描述了根据因果网络判断哪些变量是混杂因素, 哪些变量不是混杂因素的方法.

文献 [119] 基于因果网络模型提出了前门准则的可识别方法, 传统的流行病学没能意识到这个新奇的结果. 基于因果网络模型方法的弱点是, 在实际中很难得到一个已知的因果网络. 潜在结果模型的方法不需要一个已知的因果网络, 但是需要可忽略处理分配假定或者工具变量假定 [29]. 文献 [120] 将因果网络与潜在结果模型结合, 给出了判断混杂因素的综合准则, 不要求已知一个完整的因果网络,

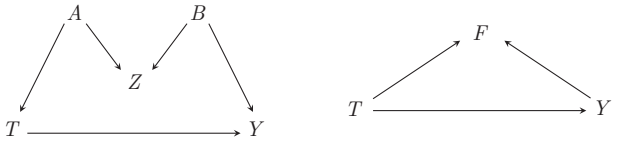

图 4　Z 和 F 都不是混杂因素

但需要一个充分大的集合其包含处理变量的父节点集合.

因果网络学习又分为因果网络的参数学习和结构学习. 参数学习是在因果网络结构已知的条件下利用数据估计参数. 根据因果网络的联合分布, 采用最大似然估计或 Bayes 方法估计条件概率 $\mathrm{pr}(x_i \mid pa_i)$. 在正态随机变量的情形, 估计给定父节点变量集合 pa_i 下 X_i 的条件正态分布的参数.

因果网络结构学习有两类方法: 基于评分的搜索方法和基于条件独立检验的方法. 20 世纪 90 年代伊始, 条件独立性检验的方法已被用于因果网络结构的发现. 文献 [121] 提出了 IC (inductive causation) 算法, 首先针对任意两个节点 X_i 和 X_j 穷尽搜索是否存在分离集 S_{ij} 使得条件独立 $X_i \perp\!\!\!\perp X_j \mid S_{ij}$ 成立; 如果存在这样的 S_{ij}, 则删去这两个节点间的边. 然后利用两个不相邻的节点 X_i 和 X_j, 如果它们的公共邻居 X_k 不包含在它们的分离集 S_{ij} 中, 则确定一个 V- 结构 $(X_i \to X_k \leftarrow X_j)$. 最后确定其他边的方向, 避免出现新的 V- 结构和有向环. 在 IC 算法中穷尽搜素分离集 S_{ij} 的计算复杂度很高, 并且对于大的分离集 S_{ij}, 其条件独立检验功效低. 为了改善 IC 算法的效率, 文献 [122] 提出了 PC (Peter and Clark) 算法. PC 算法仍是将完全图作为初始骨架图, 然后从空集开始逐步增大分离集的大小, 不断删除骨架图中的边, 使得每个结点的邻居数不断减少, 寻找两个节点的分离集限定在它们的邻居集的子集范围内, 目的是避免高维变量的条件独立检验. 后续的研究对 PC 算法不断改进, 克服了其在稳定性、潜在混杂变量处理、非线性因果关系处理、混合变量处理等方面的不足. 文献 [123] 提出了 Stable PC 算法, 通过对骨架学习及因果定向规则的修改, 降低了 PC 算法对随机变量的序列关系的敏感度, 使其在高维变量的场景中仍能获得稳定的学习效果. FCI/RFCI 算法及其变体被提出 [116, 124], 用以在未观测混杂变量和样本选择偏差存在的情形下学习因果结构, 这些方法在 PC 邻接搜索的基础上, 利用额外的条件独立性检验以处理潜在混杂变量. 文献 [125] 提出了基于核的独立性检验方式并给出条件独立零假设下的渐近分布构造方法, 以支持非线性假设下的因果关系发现. Copula PC [126] 将 PC 算法中基于相关矩阵的独立性检验更新为基于 Gauss 耦合函数相关矩阵的独立性检验, 以支持混合变量 (连续和离散变量并存) 下的因果结构学习.

针对大规模网络和多个不完全数据库, 文献 [127] 提出了网络结构的分解学习算法. 在有变量间条件独立的先验知识情形下可以不必同时观测所有的变量, 利用多个不完全变量的数据库, 首先各自学习局部网络, 然后合并为一个完整的大网络. 文献 [128] 提出了网络结构的递归学习算法, 将一个大网络结构学习分解为两个局部小网络的结构学习, 然后递归地将每个局部网络的结构学习再分解为更小网络的结构学习. 文献 [129] 提出了最小 d- 分离树的分解学习算法. 文献 [130] 提出了寻找最小分离集从道德图学习因果网络的算法. 采用分治方法实现高维因果关系, 文献 [131] 提出了一种基于 causal cut set 的变量集分割方法, 通过递归地将高维问题分解为 ANM 和 LiNGAM (linear non-Gaussian acyclic model) 等模型可以求解小规模问题, 实现了高维问题的求解. 针对 causal cut set 的计算代价大和分割过程导致误差累积等问题, 文献 [132] 提出了一种随机子问题分割、全局优化的因果序推断和冗余剔除策略的高效因果机制发现算法.

利用观测数据仅能学到一组具有相同条件独立性的网络结构, 这组网络有相同的边和相同的 V-结构, 但是它们的某些边可能有不同的方向, 这样一组网络图称为 Markov 等价类. 为了从等价类中确定哪一个网络是真网络, 需要额外的先验知识或干预数据. 文献 [133] 提出了最优干预设计的主动学习方法, 希望干预最少的变量能确定所有边的方向. 文献 [134] 探讨了利用干预试验学习 Markov 等价类的算法.

当目的是为了寻找和确定某个给定的目标变量的局部因果关系, 发现它的原因是什么和结果是什么, 或者目的是做干预决策时, 我们只需利用数据发现该目标变量的局部因果关系, 没必要学习整个网络的结构. 文献 [135] 提出了学习局部网络结构的方法, 寻找目标变量 Y 的父亲 - 儿子 - 子孙

(parents-children-descendants, PCD) 和父子 (PC) 算法, 但是他们的算法不能确定哪个是父节点, 哪个是子节点. 文献 [136] 提出了逐步构建目标变量 Y 的局部网络的方法, 称为 MB-by-MB (Markov blanket by Markov blanket) 算法. 该算法以目标变量 Y 为中心, 逐步向外扩散地构建局部网络结构. 首先寻找目标变量 Y 的 Markov 边界 MB(Y), 构建 MB(Y) 的局部网络结构; 然后再寻找 MB(Y) 中每个 X_i 的 MB(X_i), 构建 MB(X_i) 的局部网络结构; 重复这个过程直至能够区别谁是 Y 的原因, 谁是 Y 的结果为止.

另一类因果网络结构学习方法是根据评分选择最佳网络. 为每个网络赋一个评分 (如后验概率、BIC (Bayesian information criterion) 和 AIC (Akaike information criterion) 等), 搜索最佳评分的有向无环图, 常采用贪心法等启发式搜索方法. Heckerman [137] 给出了 Bayes 方法学习因果网络的方法. 在整个网络空间搜索最佳评分的网络是一个非常困难的问题. 此类方法通过定义可分解的评分准则来评价数据和网络的拟合度, 并以该准则指导最优网络结构的搜索. 当定义的评分准则满足评分等价性 (score equivalence) [138], 即等价类中的 DAG 拥有相同的分数时, 该准则可用于指导因果结构的学习. 文献 [139] 提出了两阶段的贪婪等价搜索算法 (greedy equivalence search, GES) 是该类型的一个代表性方法, 它基于观测数据, 从 DAG 空间中搜索获取真实分布的完备图 (perfect map), 但该方法尚无法处理未观测混杂变量、样本选择偏差等问题. 很多学者探索了随机模拟的搜索算法. 文献 [140] 提出了一种将约束满足、蚁群优化和模拟退火策略相结合的混合算法. 该算法首先利用阈值自调整的条件测试来动态地压缩搜索空间以加速搜索过程, 然后利用互信息修正了蚁群算法的启发函数以保证学习的求解质量, 最后通过引入模拟退火的优化调节机制改进了算法的优化效率. 文献 [141] 提出了一种基于蜂群觅食原理的因果网络结构学习新算法. 该算法通过模拟蜜蜂采集花蜜过程中雇佣蜂、观察蜂和侦察蜂三种蜂的不同觅食行为, 实现了解的局部开采和全局探测, 能够在较短时间内获得全局满意解. 文献 [142] 提出了一种基于菌群觅食原理的因果网络结构学习新算法. 该算法模拟菌群觅食过程中的趋向、复制和迁徙三种操作, 实现了寻优过程中解的开采和探测的平衡, 能有效地完成最优解的评分搜索. 文献 [143] 探索了基于群智能搜索算法, 首先从随机搜索的角度分析并概括了蚁群、蜂群和菌群三种不同搜索机理中共同的特征和不同的寻优机制, 然后用丰富的试验验证了这些不同机制各自的作用, 并对比了不同算法在有噪声数据情形下的鲁棒性. 文献 [144] 将基于群智能搜索的因果网络结构学习算法应用于脑科学中, 提出了一种基于人工免疫算法的脑效应连接网络学习方法.

将前面的条件独立检验算法和评分算法结合, 文献 [135] 提出了混合学习方法 (max-min hill-climbing, MMHC) 进行网络结构挖掘. 首先运用基于 (条件) 独立性检验的局部结构搜索算法确定因果网络的骨架, 继而利用基于 Bayes 评分准则的贪婪爬山搜索算法确定骨架中变量间的因果方向. 该方法能较好地校正基于统计独立性的方法在定向上的错误. H2PC (hybrid hybrid parents and children) 算法 [145] 也采用了类似的两阶段混合结构学习思想. MMHC 和 H2PC 方法均支持大规模随机变量上的因果结构学习.

将因果网络进行参数化, 利用结构方程模型 (structure equation model, SEM) 描述变量间的因果关系. 将结果变量 Y 与直接原因变量集合 X 和噪声项 ε 用结构方程 $Y = f(X, \varepsilon)$ 联系起来, 其中 X 与 ε 相互独立. 因果方向的可判定问题是 SEM 研究中的一项重要课题. 文献 [146, 147] 的研究表明, 当噪声项服从非 Gauss 分布或者函数方程满足非线性约束时, 由于原因变量和噪声项间的独立性仅在正确的因果方向下成立, 使得变量间的因果方向是可判定的. LiNGAM [146] 是该研究方向的一个代表性模型, 它建模连续随机变量间的因果关系, 假设变量间线性关联且噪声项服从非 Gauss 分布. 独立成分分析技术 (independent component analysis, ICA) 被用于 LiNGAM 的模型选择, 由于超参数选择

问题, ICA 算法常常陷入局部最优而无法收敛于最优解. 为此, DirectLiNGAM 算法 [148] 利用外部变量及非外部变量在其上的回归残差间的独立性信息求解变量间的因果结构, 被证明可以收敛于最优. 后续研究针对 LiNGAM 在诸多方向上进行了扩展. 例如, 文献 [149] 将潜在混杂变量及其对观测变量的影响进行建模, 并利用过完备 ICA 算法实现模型选择. 文献 [150] 提出了一种 Bayes 方法求解带潜在混杂因子的 LiNGAM 模型. 文献 [151, 152] 将 LiNGAM 进行了有环化的扩展, 并给出了模型可判定的充分条件. 文献 [153] 扩展 LiNGAM 以处理随机变量间的非线性因果关系, 并证明除个别非线性函数及数据分布外, 其模型是可判定的. 文献 [154] 对这些非线性模型进行扩展以处理潜在混杂变量. 噪声可加模型 (additive noise model, ANM) 将因果关系建模为 $Y = f(X) + \varepsilon$, 文献 [147] 的研究表明, 当函数满足非线性约束时, 该模型是可判定模型. 文献 [147] 利用目标变量与源变量在其上非线性回归的残差之间的独立性信息判定因果方向. 文献 [155] 提出了基于 Bayes 评分准则的非线性 ANM 的模型选择算法. 文献 [156] 致力于有环 ANM 的模型选择等. 针对似然函数方法存在的 Markov 等价类问题, 文献 [157] 通过将结构方程模型引入到似然函数计算框架中, 实现了似然函数方法和结构方程模型的有效结合较好地解决了 Markov 等价类问题. 文献 [158] 提出了一种结合探索性因子分析和路径分析方法推断存在隐变量情形下的因果关系, 利用因子分析得到相对各自独立的隐变量, 采用路径分析 (path analysis, PA) 算法得到观测变量之间的因果方向和因果关系, 扩展了隐变量以及它们与观察变量之间的线性因果关系.

还有一些研究将独立性检验和逻辑推理结合, 以解决 PC 类算法的稳健性问题 (不正确的独立性检验会导致连锁的定向错误). 该类方法的另一个优势在于易集成多种类别的先验知识, 同时易于处理未观测的混杂因素及数据选择偏差. 文献 [159] 将变量间的 (条件) 独立信息转换成逻辑命题, 并给出了相应的逻辑推理算法以识别部分祖先图 (partial ancestral graph, PAG). 文献 [160] 提出使用一阶逻辑编码随机变量间的条件独立关系, 将因果关系发现问题转换成骨干变量求解问题, 并利用 Boolean satisfiability (SAT) 处理器识别因果结构. 文献 [161] 的工作更近一步定义了更多的逻辑项及规则对控制试验数据、非同源数据等信息进行编码及推理, 实现了更泛化的因果结构学习.

针对高维稀疏图模型, 文献 [162] 构造了图空间上的可逆 Markov 链, 实现了高维稀疏图的高效随机抽样方法, 可有效地应用于图模型结构学习的 Bayes 方法. 文献 [163] 通过一种递归算法解决了等价类中图模型结构数量的计数问题, 应用于高维稀疏图模型等价类, 快速计数一个等价类中包含的图模型个数, 对于分析图模型方法的复杂性和因果推断具有重要的作用. 将因果网络结构学习应用于大数据中, 文献 [164] 针对数据的海量、分布式和动态变化特征, 扩展了用于因果网络学习的评分搜索算法, 提出了基于 MapReduce 编程模型的因果网络并行学习和增量维护方法. 文献 [165] 从海量的社交用户行为交互中构建描述用户之间依赖关系的因果网络, 以及基于 MapReduce 编程模型的大规模因果网络概率推理算法, 为社交网应用中关联分析和相似搜索提供支撑技术. 文献 [166] 引入因果度量挖掘药物相互作用导致的不良反应机制. 文献 [167] 通过基于因果机制的用户行为序列分析, 发现了同质和反向影响等社交网络行为的隐藏原因.

近年来, 将在因果网络的因果作用可识别性问题与因果网络的学习问题结合, 很多学者探讨了数据驱动的因果推断方法. 这类方法首先利用数据学习因果网络的 Markov 等价类, 然后识别等价类中每个网络的因果作用, 最后得到所有可能的因果作用的集合或者上下界. 对于高维因果网络, 一个 Markov 等价类可能包括大量的 Markov 性等价的因果网络, 枚举所有可能的因果网络是一个困难的问题, 而且不同的因果网络可能有相等的因果作用. 给定表示 Markov 等价类的本质图, 文献 [168] 提出了从本质图中枚举所有可能的因果作用的局部算法, 得到所有因果作用的集合和上下界. 针对多个干预的情形, 文献 [169] 进一步提出了估计多处理对结果变量的联合因果作用的上下界的局部算法. 针

对线性结构方程模型, 文献 [170] 提出了采用数据学习因果网络的等价类, 允许潜在混杂因素的存在, 然后针对等价类中每个模型根据所有可能的因果作用的界. 文献 [171] 提出了数据驱动的混杂因素选择, 给出了利用数据学习因果网络结构与因果作用估计相结合的方法.

因果网络是多变量之间因果关系的重要形式化方法, 被广泛应用于各领域的科学研究, 包括基因调控和脑神经调控等生命科学领域. 但是, 因果网络的结构学习需要相当大的样本和有关忠实性的假定; 当存在有隐变量、因果反馈等情形时, 仍有待于方法论上的突破. 从观测数据学习得到因果网络的 Markov 等价类仍然包含众多可能的因果网络, 如何利用专业背景知识和合理假定更细致地发现变量之间的因果关系仍有待于研究. 另外, 采用因果网络形式化地描述多变量之间的因果关系和外部干预也有待于深入探讨.

7 结束语

研究因果关系一直是人类探索世界的主题. 挖掘因果关系的科学方法, 对各个科学研究领域都有普适性. 现今的生命科学、信息科学、社会科学和经济金融等领域都迫切需要这些方法. 在哲学史上, 远至亚里士多德的 “四因说”, 近至 “穆勒五法” 和 “休谟问题”, 都对因果概念做了深入透彻的论述. 关于因果关系的哲学思想发展史, 参见文献 [172]. 但是, 这些学说对现代科学研究中发现因果关系的方法上的指导作用很有限. 在近一个世纪, 随着统计学的快速发展, 因果推断的统计方法在诸多科学领域如流行病学、生物医药、社会学和经济学取得了辉煌成就, 并在其他学科中展现出巨大潜力. 本文介绍的潜在结果模型和因果网络模型, 即是在评价因果作用和发现因果关系中最成熟和应用最广泛的统计方法. 这些方法的成功, 得益于定量化数据记录对各个学科的普适性, 以及统计学以数据为分析对象的特点.

不可否认, 因果推断必需的一些基本假定在实际中无法用数据完全验证, 例如, 可忽略性假定和 SUTVA 假定; 这些假定是否成立需要根据专业知识或者先验知识来判断. 在实际研究中做因果推断, 要充分理解这些假定的含义才能根据专业知识判断. 为了得到可靠的因果推论, 需要考查不同的假定对因果推断的影响, 进行有关假定的敏感性分析 (sensitivity analysis), 参见文献 [90]. 本文未涉及的因果推断中其他重要的问题, 如中介分析 [173, 174]、主分层作用 [106, 175]、不依从性问题 [73, 176] 和干涉问题 [13, 16] 等, 在生物医学、流行病学和社会经济学中都有重要作用.

观察性研究中的混杂因素和不依从性的问题, 在现代大数据研究中不可避免, 也揭示了大数据研究潜在的缺陷; 工具变量和阴性对照变量方法则启发我们在研究设计和数据收集过程中就应记录一些辅助数据, 而不应只关注目标变量. 替代指标悖论则揭示了利用多个统计结论进行联合推理的困难. 文献 [177] 探讨了密度关联性、分布关联性、期望关联性和线性相关性等关联度量的可传递性. 替代指标悖论启发我们, 从大数据的多源数据库得到的众多统计结论存在传递性的问题. 因此, 对于多源数据库需要先融合数据, 然后再对融合的数据进行分析; 而不应当从孤立的数据库得到各自的结论, 然后用这些结论进行推理.

在大数据时代, 数据的收集和分析在各个学科和研究领域都变得越来越重要, 而根据数据推断因果作用和寻找因果关系将成为推动各个学科和领域发展的重要动力, 因果推断方法必将大展神通. 如果把 Pearson 和 Fisher 时代比作统计学的 Newton 时代, 我们期待着大数据时代将会出现统计学的 Einstein 式人物.

致谢 感谢主编以及两位匿名审稿人的宝贵意见.

参考文献

1 Holland P W. Statistics and causal inference. J Amer Statist Assoc, 1986, 81: 945–960
2 Pearl J. The art and science of cause and effect. In: Causality: Models, Reasoning, and Inference, 2nd ed. New York: Cambridge University Press, 2009, 401–428
3 Lauritzen S L. Discussion on causality. Scand J Statist, 2004, 31: 189–193
4 Guyon I, Aliferis C, Cooper G F, et al. Design and analysis of the causation and prediction challenge. Proceed J Mach Learn Res, 2008, 3: 1–33
5 Spirtes P. Introduction to causal inference. J Mach Learn Res, 2010, 11: 1643–1662
6 Heckman J J. Econometric causality. Int Statist Rev, 2008, 76: 1–27
7 Granger C W J. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 1969, 37: 424–438
8 Neyman J. On the application of probability theory to agricultural experiments. Trans Statist Sci, 1923, 5: 465–480
9 Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol, 1974, 66: 688–701
10 Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann, 1988
11 Pearl J. Causality: Models, Reasoning, and Inference, 2nd ed. New York: Cambridge University Press, 2009
12 Rubin D B. Randomization analysis of experimental data: The Fisher randomization test comment. J Amer Statist Assoc, 1980, 75: 591–593
13 Hudgens M G, Halloran M E. Toward causal inference with interference. J Amer Statist Assoc, 2008, 103: 832–842
14 Sobel M E. What do randomized studies of housing mobility demonstrate? J Amer Statist Assoc, 2006, 101: 1398–1407
15 Tchetgen E J T, Van der Weele T J. On causal inference in the presence of interference. Stat Methods Med Res, 2012, 21: 55–75
16 Liu L, Hudgens M G. Large sample randomization inference of causal effects in the presence of interference. J Amer Statist Assoc, 2014, 109: 288–301
17 Eckles D, Karrer B, Ugander J. Design and analysis of experiments in networks: Reducing bias from interference. J Causal Inference, 2017, 5: doi: 10.1515/jci-2015-0021
18 Athey S, Eckles D, Imbens G W. Exact p-values for network interference. J Amer Statist Assoc, 2018, 113: 230–240
19 Rubin D B. Bayesian inference for causal effects: The role of randomization. Ann Statist, 1978, 6: 34–58
20 Murphy S A. Optimal dynamic treatment regimes. J R Stat Soc Ser B Stat Methodol, 2003, 65: 331–355
21 Su X, Kang J, Fan J, et al. Facilitating score and causal inference trees for large observational studies. J Mach Learn Res, 2012, 13: 2955–2994
22 Kleinberg S, Hripcsak G. A review of causal inference for biomedical informatics. J Biol Med Inf, 2011, 44: 1102–1112
23 Chakraborty B, Moodie E. Statistical Methods for Dynamic Treatment Regimes. New York: Springer, 2013
24 Fisher R A. Design of Experiments. Edinburgh: Oliver and Boyd, 1935
25 Cochran W G, Chambers S P. The planning of observational studies of human populations. J Roy Statist Soc Ser A, 1965, 128: 234–266
26 Yule G U. Notes on the theory of association of attributes in statistics. Biometrika, 1903, 2: 121–134
27 Simpson E H. The interpretation of interaction in contingency tables. J R Stat Soc Ser B Stat Methodol, 1951, 13: 238–241
28 Bickel P J, Hammel E A, O'Connell J W. Sex bias in graduate admissions: Data from Berkeley. Science, 1975, 187: 398–404
29 Pearl J, Robins J M, Greenland S. Confounding and collapsibility in causal inference. Statist Sci, 1999, 14: 29–46
30 Miettinen O S, Cook E F. Confounding: Essence and detection. Amer J Epidemiol, 1981, 114: 593–603
31 Kleinbaum D G, Kupper L L, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. New York: John Wiley & Sons, 1982
32 Greenland S, Robins J M. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol, 1986, 15: 413–419
33 Geng Z, Guo J, Fung W K. Criteria for confounders in epidemiological studies. J R Stat Soc Ser B Stat Methodol, 2002, 64: 3–15
34 Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70: 41–55
35 Cochran W G, Rubin D B. Controlling bias in observational studies: A review. Sankhya Ser A, 1973, 35: 417–446
36 Rubin D B. Matching to remove bias in observational studies. Biometrics, 1973, 29: 159–183

37 Abadie A, Imbens G W. Matching on the estimated propensity score. Econometrica, 2016, 84: 781–807
38 Lee B K, Lessler J, Stuart E A. Improving propensity score weighting using machine learning. Stat Med, 2010, 29: 337–346
39 Abadie A, Imbens G W. Large sample properties of matching estimators for average treatment effects. Econometrica, 2006, 74: 235–267
40 Austin P C. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med, 2008, 27: 2037–2049
41 Austin P C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behaval Res, 2011, 46: 399–424
42 Stuart E A. Matching methods for causal inference: A review and a look forward. Statist Sci, 2010, 25: 1–21
43 Hansen L P. Large sample properties of generalized method of moments estimators. Econometrica, 1982, 50: 1029–1054
44 Hall A R. Generalized Method of Moments. Oxford: Oxford University Press, 2005
45 Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe. J Amer Statist Assoc, 1952, 47: 663–685
46 Robins J M, Rotnitzky A, Zhao L P. Estimation of regression coefficients when some regressors are not always observed. J Amer Statist Assoc, 1994, 89: 846–866
47 Rotnitzky A, Robins J M, Scharfstein D O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. J Amer Statist Assoc, 1998, 93: 1321–1339
48 Scharfstein D O, Rotnitzky A, Robins J M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Amer Statist Assoc, 1999, 94: 1096–1120
49 Robins J M, Rotnitzky A. Comment on the bickel and kwon article, “on double robustness”. Statist Sinica, 2001, 11: 920–936
50 Van der Laan M J, Robins J M. Unified Methods for Censored Longitudinal Data and Causality. New York: Springer, 2003
51 Tsiatis A. Semiparametric Theory and Missing Data. New York: Springer, 2006
52 Lipsitz S R, Ibrahim J G, Zhao L P. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. J Amer Statist Assoc, 1999, 94: 1147–1160
53 Robins J M, Rotnitzky A, van der Laan M. On profile likelihood: Comment. J Amer Statist Assoc, 2000, 95: 477–482
54 Lunceford J K, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statist Med, 2004, 23: 2937–2960
55 Bang H, Robins J M. Doubly robust estimation in missing data and causal inference models. Biometrics, 2005, 61: 962–973
56 Vansteelandt S, Rotnitzky A, Robins J. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. Biometrika, 2007, 94: 841–860
57 Miao W, Tchetgen Tchetgen E J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. Biometrika, 2016, 103: 475–482
58 Liu L, Miao W, Sun B, et al. Doubly robust estimation of a marginal average effect of treatment on the treated with an instrumental variable. Statist Sinica, 2018, in press
59 Sun B, Liu L, Miao W, et al. Semiparametric estimation with data missing not at random using an instrumental variable. Statist Sinica, 2018, 28: 1965–1983
60 Kang J D Y, Schafer J L. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist Sci, 2007, 22: 523–539
61 Robins J, Sued M, Lei-Gomez Q, et al. Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. Statist Sci, 2007, 22: 544–559
62 Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. Biometrika, 2010, 97: 661–682
63 Vermeulen K, Vansteelandt S. Bias-reduced doubly robust estimation. J Amer Statist Assoc, 2015, 110: 1024–1036
64 Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Amer Statist Assoc, 1997, 92: 1171–1176
65 Manski C F. Nonparametric bounds on treatment effects. Amer Econ Rev, 1990, 80: 319–323
66 Richardson T S, Evans R J, Robins J M. Transparent parameterizations of models for potential outcomes. Bayesian Stat, 2011, 9: 569–610
67 Hernán M A, Robins J M. Causal Inference. Boca Raton: Chapman & Hall, 2018
68 Wright P G. Tariff on Animal and Vegetable Oils. New York: Macmillan, 1928
69 Goldberger A S. Structural equation methods in the social sciences. Econometrica, 1972, 40: 979–1001

70 Wooldridge J M. Econometric Analysis of Cross Section and Panel Data. Cambridge: MIT press, 2010
71 Hernán M A, Robins J M. Instruments for causal inference: An epidemiologist's dream? Epidemiology, 2006, 17: 360–372
72 Imbens G W, Angrist J D. Identification and estimation of local average treatment effects. Econometrica, 1994, 62: 467–475
73 Angrist J D, Imbens G W, Rubin D B. Identification of causal effects using instrumental variables. J Amer Statist Assoc, 1996, 91: 444–455
74 Newey W K, Powell J L. Instrumental variable estimation of nonparametric models. Econometrica, 2003, 71: 1565–1578
75 Ai C, Chen X. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica, 2003, 71: 1795–1843
76 Clarke P S, Windmeijer F. Instrumental variable estimators for binary outcomes. J Amer Statist Assoc, 2012, 107: 1638–1652
77 Vansteelandt S, Bowden J, Babanezhad M, et al. On instrumental variables estimation of causal odds ratios. Statist Sci, 2011, 26: 403–422
78 Burgess S, Small D S, Thompson S G. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res, 2017, 26: 2333–2355
79 Bound J, Jaeger D A, Baker R M. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Amer Statist Assoc, 1995, 90: 443–450
80 Stock J H, Wright J H, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econom Statist, 2002, 20: 518–529
81 Baiocchi M, Cheng J, Small D S. Instrumental variable methods for causal inference. Statist Med, 2014, 33: 2297–2340
82 Hill A B. The environment and disease: Association or causation? Proc Roy Soc Med, 1965, 58: 295–300
83 Berkson J. Smoking and lung cancer: Some observations on two recent reports. J Amer Statist Assoc, 1958, 53: 28–38
84 Yerushalmy J, Palmer C E. On the methodology of investigations of etiologic factors in chronic diseases. J Chronic Dis, 1959, 10: 27–40
85 Weiss N S. Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology, 2002, 13: 6–8
86 Rosenbaum P R. The role of known effects in observational studies. Biometrics, 1989, 45: 557–569
87 Flanders W D, Klein M, Darrow L A, et al. A method for detection of residual confounding in time-series and other observational studies. Epidemiology, 2011, 22: 59–67
88 Trichopoulos D, Zavitsanos X, Katsouyanni K, et al. Psychological stress and fatal heart attack: The athens (1981) Earthquake natural experiment. Lancet, 1983, 321: 441–444
89 Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: A tool for detecting confounding and bias in observational studies. Epidemiology, 2010, 21: 383–388
90 Rosenbaum P R. Observational Studies, 2nd ed. New York: Springer, 2002
91 Schuemie M J, Ryan P B, DuMouchel W, et al. Interpreting observational studies: Why empirical calibration is needed to correct p-values. Statist Med, 2014, 33: 209–218
92 Gagnon-Bartsch J A, Speed T P. Using control genes to correct for unwanted variation in microarray data. Biostatistics, 2012, 13: 539–552
93 Sofer T, Richardson D B, Colicino E, et al. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. Statist Sci, 2016, 31: 348–361
94 Miao W, Geng Z, Tchetgen Tchetgen E. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika, 2018, 105: 987–993
95 Miao W, Tchetgen Tchetgen E. Invited commentary: Bias attenuation and identification of causal effects with multiple negative controls. Amer J Epidemiol, 2017, 185: 950–953
96 Miao W, Tchetgen Tchetgen E. A confounding bridge approach for double negative control inference on causal effects. ArXiv:1808.04945, 2018
97 D'Haultfoeuille X. On the completeness condition in nonparametric instrumental problems. Econometric Theory, 2011, 27: 460–471
98 Darolles S, Fan Y, Florens J P, et al. Nonparametric instrumental regression. Econometrica, 2011, 79: 1541–1565
99 Chen X H, Chernozhukov V, Lee S, et al. Local identification of nonparametric and semiparametric models. Econometrica, 2014, 82: 785–809
100 Andrews D W K. Examples of L^2-complete and boundedly-complete distributions. J Econometrics, 2017, 199: 213–

220
101 Fleming T R, Demets D L. Surrogate end points in clinical trials: Are we being misled? Ann Internat Med, 1996, 125: 605–613
102 Baker S. Surrogate endpoints: Wishful thinking or reality? J Natl Cancer I, 2006, 98: 502–503
103 Manns B, Owen W F, Winkelmayer W C, et al. Surrogate markers in clinical studies: Problems solved or created? Amer J Kidney Diseases, 2006, 48: 159–166
104 Alonso A, Molenberghs G. Surrogate end points: Hopes and perils. Expert Rev Pharmacoeconomics Outcomes Res, 2008, 8: 255–259
105 Prentice R L. Surrogate endpoints in clinical trials: Definition and operational criteria. Statist Med, 1989, 8: 431–440
106 Frangakis C E, Rubin D B. Principal stratification in causal inference. Biometrics, 2002, 58: 21–29
107 Chen H, Geng Z, Jia J. Criteria for surrogate end points. J R Stat Soc Ser B Stat Methodol, 2007, 69: 919–932
108 Moore T. Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster. New York: Simon & Schuster, 1995
109 VanderWeele T J. Surrogate measures and consistent surrogates. Biometrics, 2013, 69: 561–565
110 Pearl J. Is scientific knowledge useful for policy analysis? J Causal Inference, 2014, 2: 109–112
111 Ju C, Geng Z. Criteria for surrogate end points based on causal distributions. J R Stat Soc Ser B Stat Methodol, 2010, 72: 129–142
112 Wu Z, He P, Geng Z. Sufficient conditions for concluding surrogacy based on observed data. Statist Med, 2011, 30: 2422–2434
113 Luo P, Cai Z, Geng Z. Criteria for multiple surrogates. Statist Sinica, 2018, in press
114 Lauritzen S L, Spiegelhalter D J. Local computations with probabilities on graphical structures and their application to expert systems. J R Stat Soc Ser B Stat Methodol, 1988, 50: 157–224
115 Spiegelhalter D J, Dawid A P, Lauritzen S L, et al. Bayesian analysis in expert systems. Statist Sci, 2011, 9: 219–83
116 Spirtes P, Glymour C, Sces R. Causation, Prediction, and Search, 2nd ed. Cambridge: MIT Press, 2000
117 Greenland S, Pearl J. Adjustments and their consequences-collapsibility analysis using graphical models. Int Statist Rev, 2011, 79: 401–426
118 Greenland S, Pearl J, Robins J M. Causal diagrams for epidemiologic research. Epidemiology, 1999, 10: 37–48
119 Pearl J. Causal diagrams for empirical research. Biometrika, 1995, 82: 669–688
120 Geng Z, Li G. Conditions for non-confounding and collapsibility without knowledge of completely constructed causal diagrams. Scand J Statist, 2002, 29: 169–181
121 Verma T, Pearl J. Equivalence and synthesis of causal models. In: Proceedings of the 6th Conference on Uncertainy in Artificial Intelligence. Amsterdam: Elsevier, 1990, 255–270
122 Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev, 1991, 9: 62–72
123 Colombo D, Maathuis M H. Order-independent constraint-based causal structure learning. J Mach Learn Res, 2014, 15: 3921–3962
124 Spirtes P, Meek C, Richardson T S. An algorithm for causal inference in the presence of latent variables and selection bias. In: Computation, Causation, and Discovery. New Orleans: AAAI Press, 1999, 211–252
125 Zhang K, Peters J, Janzing D, et al. Kernel-based conditional independence test and application in causal discoveyr. Comput Sci, 2012, 8: 895–907
126 Cui R, Groot P, Heskes T. Copula PC algorithm for causal discovery from mixed data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 9852. Cham: Springer, 2016, 377–392
127 Xie X, Geng Z, Zhao Q. Decomposition of structural learning about directed acyclic graphs. Artificial Intelligence, 2006, 170: 422–439
128 Xie X, Geng Z. A recursive method for structural learning of directed acyclic graphs. J Mach Learn Res, 2008, 9: 459–483
129 Liu B, Guo J, Jing B Y. A note on minimal d-separation trees for structural learning. Artificial Intelligence, 2010, 174: 442–448
130 Xu P F, Guo J, Tang M L. Structural learning for bayesian networks by testing complete separators in prime blocks. Comput Statist Data Anal, 2011, 55: 3135–3147
131 Cai R, Zhang Z, Hao Z. Sada: A general framework to support robust causation discovery. In: Proceedings of the 30th International Conference on Machine Learning. Atlanta: International Machine Learning Society, 2013, 208–216
132 Cai R C, Zhang Z J, Hao Z F, et al. Sophisticated merging over random partitions: A scalable and robust causal discovery approach. IEEE Trans Neural Netw Learn Syst, 2018, 29: 3623–3635
133 He Y B, Geng Z. Active learning of causal networks with intervention experiments and optimal designs. J Mach

Learn Res, 2008, 9: 2523–2547
134 Hauser A, Bühlmann P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. J Mach Learn Res, 2012, 13: 2409–2464
135 Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn, 2006, 65: 31–78
136 Wang C Z, Zhou Y, Zhao Q, et al. Discovering and orienting the edges connected to a target variable in a DAG via a sequential local learning approach. Comput Statist Data Anal, 2014, 77: 252–266
137 Heckerman D. A tutorial on learning with Bayesian networks. In: Innovations in Bayesian Networks. Studies in Computational Intelligence, vol 156. Berlin-Heidelberg: Springer, 2008
138 Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: The combination of knowledge and statistical data. Mach Learn, 1995, 20: 197–243
139 Chickering D M. Optimal structure identification with greedy search. Mach Learn, 2002, 3: 507–554
140 Ji J, Hu R, Zhang H, et al. A hybrid method for learning Bayesian networks based on ant colony optimization. Appl Soft Comput, 2011, 11: 3373–3384
141 Ji J, Wei H, Liu C. An artificial bee colony algorithm for learning Bayesian networks. Soft Comput, 2013, 17: 983–994
142 Yang C, Ji J, Liu J, et al. Structural learning of Bayesian networks by bacterial foraging optimization. Internat J Approx Reason, 2016, 69: 147–167
143 Ji J, Yang C, Liu J, et al. A comparative study on swarm intelligence for structure learning of Bayesian networks. Soft Comput, 2017, 21: 6713–6738
144 Ji J, Liu J, Liang P, et al. Learning Effective connectivity network structure from fMRI data based on artificial immune algorithm. PLoS ONE, 2016, 11: e0152600
145 Gasse M, Aussem A, Elghazel H. A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. Expert Syst Appl, 2014, 41: 6755–6772
146 Shimizu S, Hoyer P O, Hyvarinen A, et al. A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res, 2006, 7: 2003–2030
147 Hoyer P O, Janzing D, Mooij J M, et al. Nonlinear causal discovery with additive noise models. In: Proceedings of the 21th Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2009, 689–696
148 Shimizu S, Inazumi T, Sogawa Y, et al. Directlingam: A direct method for learning a linear non-gaussian structural equation model. J Mach Learn Res, 2011, 12: 1225–1248
149 Hoyer P O, Shimizu S, Kerminen A J, et al. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. Internat J Approx Reason, 2008, 49: 362–378
150 Henao R, Winther O. Sparse linear identifiable multivariate modeling. J Mach Learn Res, 2011, 12: 863–905
151 Lacerda G, Spirtes P, Ramsey J, et al. Discovering cyclic causal models by independent components analysis. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2008, 366–374
152 Hyvarinen A, Smith S M. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. J Mach Learn Res, 2013, 14: 111–152
153 Zhang K, Hyvarinen A. On the identifiability of the post-nonlinear causal model. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2009, 647–655
154 Zhang K, Schölkopf B, Janzing D, et al. Invariate Gaussian process latent variable models and appliation in caual discovery. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2010, 717–724
155 Bühlmann P, Peters J, Ernest J. Cam: Causal additive models, high-dimensional order search and penalized regression. Ann Statist, 2013, 42: 2526–2556
156 Mooij J M, Janzing D, Heskes T, et al. On causal discovery with cyclic additive noise models. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2011, 639–647
157 Cai R C, Qiao J, Zhang Z J, et al. Self: Structural equational embedded likelihood framework for causality discovery. In: AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018, 1787–1794
158 Fei N, Yang Y. Estimating linear causality in the presence of latent variables. Cluster Comput, 2017, 20: 1025–1033
159 Claassen T, Heskes T. A logical characterization of constraint-based causal discovery. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2011, 135–144
160 Hyttinen A, Hoyer P O, Eberhardt F, et al. Discovering cyclic causal models with latent variables: A general sat-based procedure. In: Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2013, 301–310
161 Borboudakis G, Tsamardinos I. Towards robust and versatile causal discovery for business applications. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:

Amer Math Soc, 2016, 1435–1444
162 He Y, Jia J, Yu B. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. Ann Statist, 2013, 41: 1742–1779
163 He Y, Jia J, Yu B. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. J Mach Learn Res, 2015, 16: 2589–2609
164 Yue K, Fang Q, Wang X, et al. A parallel and incremental approach for data-intensive learning of bayesian networks. IEEE Trans Cybern, 2015, 45: 2890–2904
165 Yue K, Wu H, Fu X, et al. A data-intensive approach for discovering user similarities in social behavioral interactions based on the bayesian network. Neurocomputing, 2017, 219: 364–375
166 Cai R, Liu M, Hu Y, et al. Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artificial Intelligence Med, 2017, 76: 7–15
167 Cai R, Zhang Z, Hao Z, et al. Understanding social causalities behind human action sequences. IEEE Trans Neural Netw Learn Syst, 2017, 28: 1801–1813
168 Maathuis M H, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. Ann Statist, 2009, 37: 3133–3164
169 Nandy P, Maathuis M H, Richardson T S. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. Ann Statist, 2017, 45: 647–674
170 Malinsky D, Spirtes P. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. Internat J Approx Reason, 2017, 88: 371–384
171 Häggström J. Data-driven confounder selection via Markov and Bayesian networks. Biometrics, 2018, 74: 389–398
172 Beebee H, Hitchcock C, Menzies P. The Oxford Handbook of Causation. Oxford: Oxford University Press, 2009
173 Van der Weele T. Explanation in Causal Inference: Methods for Mediation and Interaction. New York: Oxford University Press, 2015
174 Shan N, Guo J. Bounds on average controlled direct effects with an unobserved response variable. J Syst Sci Complex, 2011, 24: 1154–1164
175 Shan N, Dong X, Xu P, et al. Sharp bounds on survivor average causal effects when the outcome is binary and truncated by death. ACM Trans Intell Syst Tech, 2015, 7: 1–11
176 Chen H, Geng Z, Zhou X H. Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. Biometrics, 2009, 65: 675–682
177 Jiang Z C, Ding P, Geng Z. Qualitative evaluation of associations and validation of surrogacy by association transitivity. Statist Sinica, 2015, 25: 1065–1079

Statistical approaches for causal inference

Wang Miao, Chunchen Liu & Zhi Geng

Abstract Causal inference is a permanent challenge topic in statistics, data science, and many other scientific fields. In this paper, we give an overview of statistical methods for causal inference. There are two main frameworks of causal inference: the potential outcome model and the causal network model. The potential outcome framework is used to evaluate causal effects of a known treatment or exposure variable on a given response or outcome variable. We review several commonly-used approaches in this framework for causal effect evaluation. The causal network framework is used to depict causal relationships among variables and the data generation mechanism in complex systems. We review two main approaches for structural learning: the constraint-based method and the score-based method. In the recent years, the evaluation of causal effects and the structural learning of causal networks are combined together. At the first stage, the hybrid approach learns a Markov equivalent class of causal networks from observed data; then at the second stage, it evaluates the causal effect for each causal network in the class; it also obtains a set of causal effects. The current frameworks of causal inference still have various demerits and disadvantages. We discuss these challenges and possible solutions in modern big data studies.

Keywords **causal effect, causal network, confounder, potential outcome, surrogate, statistical inference, directed acyclic graph**

MSC(2010) **62A01, 68T30**

doi: 10.1360/N012018-00055