

提名项目：因果推断的统计理论与方法

提名者：中国数学会

提名意见：

因果关系因为难以量化，因此曾经是数学的禁区。近几十年来，随着因果量化模型的出现，因果推断在统计学、计量经济学、人工智能等领域逐渐成为显学，大量的研究者投入其中。项目申请团队是国内统计因果推断方向的先驱，早在上世纪 90 年代初就很有远见地投入其中，在因果推断方向取得一系列重要的成就，具有世界级的影响力。该项目的主要贡献包括：提出该领域核心概念‘混杂因素’的严格量化定义和判断准则，从而给出了避免 Yule-Simpson 悖论现象发生的条件；创造性的提出‘替代指标悖论’，引起学界热议和大量后续研究；提出在非随机缺失和不依从数据情形下识别和评价临床试验中因果作用的方法，被国际同行评价为对随机化试验分析做出了重要贡献；提出了因果网络结构学习的分治方法，将大规模网络的学习问题分解为小网络递归学习问题，及干预设计的主动学习方法。

该项目的一系列成果发表在统计学、人工智能和机器学习的国际最权威刊物上，得到了包括图灵奖得主、美国科学院院士、英国皇家科学院院士、COPSS 奖获得者等在内的同行专家的高度评价。在国际工业数学大会 ICIAM 2011 上做 1 小时大会主旨报告。两位完成人获国家杰出青年基金（其中一位还是长江学者），两位完成人为统计学科评议组成员（其中一位还是召集人），在国家统计学学科建设中发挥了重要作用。

鉴于该项目的创新性、重要性和国际影响，中国数学会极力推荐该项目申报 2020 年度国家自然科学奖，希望专家给予支持。

提名该项目为国家自然科学奖 二 等奖

项目简介：

本项目属数理统计领域，具体对象为统计因果推断。因果关系几乎是所有科学研究的最终目标，但一直以来研究因果却缺失描述因果的数学语言。由于相关不等于因果，由此导致了很多悖论的产生，比如著名的 Yule-Simpson 悖论就描述了“变量之间的相关关系可以被第三个变量（混杂因素）逆转其正负符号”这一现象。近期以来，先后出现了两种主要的因果推断模型：一是潜在结果模型，二是因果网络模型。本团队是国内最早进入这一领域的研究团队，持续研究了近 30 年，取得的主要成果如下：

1. 针对 Yule-Simpson 悖论，本项目为“混杂因素”这一因果推断的最基础概念

提出了严格的量化定义。与传统的混杂因素定义和判定条件的根本区别是，传统的定义和条件需要处理分配可忽略性假定，这个假定本质上是不可经验论证的。而本项目不需要这个假定，给出一系列判断混杂因素的准则和充分必要条件及避免虚假相关的条件。

2. 在因果推断中，人们常常想当然地认为因果作用满足传递性。为此，替代指标 (Surrogate) 或生物标记物常用于生物医学等科学研究中。目前确定替代指标的三个主要准则是由美国科学院院士 Prentice、美国科学院院士 Rubin、丹麦皇家科学院院士 Lauritzen 先后提出的。本项目创造性地提出“替代指标悖论”：治疗对替代指标（如纠正心律失常）有正因果作用，替代指标对终点指标（寿命）有正因果作用，但是治疗可能会对终点指标有负因果作用，指出因果作用一般不满足传递性。并论述了以上三位院士提出的准则都不能避免替代指标悖论发生。本项目进一步给出了一系列避免悖论的判定条件。
3. 随机化试验是确定因果效应的‘金标准’，但随机化试验中常常出现不依从性和非随机缺失两种现象，非常难以处理。国际同行评价本研究成果‘对这两个复杂问题的解决做出了重要贡献’。本研究给出了新的非随机缺失模型的识别性准则，指出原本以为可识别的非随机缺失数据机制实际上是不可识别的，扩展了诺贝尔奖获得者 Heckman 的选择模型的可识别性结果。一篇代表性论文被主编推荐为年度唯一讨论论文。
4. 将因果推断应用于人工智能研究，提出了因果网络的各种结构学习方法（分解学习、主动学习、局部学习）。利用分治策略，将大规模网络的学习分解为小网络学习，采用递归分解进行局部学习；提出判断因果关系的主动学习方法，利用最少的干预确定因果方向。

本项目的主要成果发表在国际最权威的统计学期刊 JRSS.B、JASA 和国际最权威的人工智能与机器学习期刊 AIJ、JMLR 等上。国际统计大奖获得者、统计 COPSS 奖获得者、图灵奖获得者、数位美国、英国、丹麦的科学院院士多次引用过这里的 8 篇代表性论文。

项目完成人中有两位 ISI 推选会员、曾任美国统计学会杂志 JASA、国际统计计算学会杂志 CSDA、泛化统计学会杂志 Stat Sinica 等编委，曾应邀在 2011 年国际工业与应用数学大会 (ICIAM 2011) 做 1 小时大会主旨报告、亚太国际会议 IMS-APRM 2012 做 45 分钟特邀报告、国际会议 CSPS-IMS 2005 做 55 分钟特别讲座、日本统计计算大会做 1 小时特别讲座等。受国内外刊物主编之邀撰写因果推断综述论文多篇。

代表性论文（专著）目录（不超过 8 篇）

序号	论文（专著） 名称/刊名 /作者	年卷页码 (xx 年 xx 卷 xx 页)	发表时间 (年月 日)	通讯 作者（含 共同）	第一 作者（含 共同）	国内作者	他引总 次数	检索数 据库	论文署名 单位是否 包含国外 单位
1	Criteria for surrogate end points / J. Royal Statist. Soc. Ser. B / Chen, H; Geng, Z; Jia, JZ	2007 年 69 卷 919-932	2007 年 04 月 01 日	Geng, Zhi	Chen, Hua	陈华, 耿直, 贾金柱	19	SCI-E	否
2	Collapsibility of logistic regression coefficients. / J. Royal Statist. Soc. Ser B / Guo, JH; Geng, Z	1995 年 57 卷 263-267	1995 年 01 月 01 日	Geng, Zhi	Guo, Jianhua	郭建华, 耿直	25	SCI-E	否
3	A recursive method for structural learning of directed acyclic graphs. / J. Machine Learning Research / Xie, XC; Geng, Z	2008 年 9 卷 459-483	2008 年 08 月 03 日	Geng, Zhi	Xie, Xianchao	谢先超, 耿直	21	SCI-E	否
4	Identifiability of normal and normal mixture models with nonignorable missing data. / J. Am. Statist. Asso. / Miao, W; Ding, P; Geng, Z	2016 年 111 卷 1673-1683	2016 年 12 月 01 日	Ding, Peng	Miao, Wang	苗旺, 耿直	8	SCI-E	是
5	Criteria for surrogate endpoints based on causal distributions. / J. Royal Statist. Soc. Ser B / Ju, C; Geng, Z	2010 年 72 卷 129-142	2010 年 03 月 01 日	Geng, Zhi	Ju, Chuan	鞠川, 耿直	12	SCI-E	否

6	Criteria for confounders in epidemiological studies. / J. Royal Statist. Soc Ser B/ Geng, Z; Guo, JH; Fung, WK	2002 年 64 卷 3-15	2002 年 01 月 01 日	Geng, Zhi	Geng, Zhi	耿直, 郭建华, 冯荣锦	10	SCI-E	否
7	Principal causal effect identification and surrogate endpoint evaluation by multiple trials. / J Royal Statist. Soc. B./ Jiang, ZC; Ding, P; Geng, Z	2016 年 78 卷 829-848	2016 年 09 月 01 日	Ding, Peng	Jiang, Zhichao	蒋智超, 耿直	3	SCI-E	是
8	Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing-data (with discussion)/ Biometrics/ Chen, H; Geng, Z; Zhou, XH	2009年65卷 675-682	2009年06月 01 日	Zhou, Xiaohua	Chen, Hua	陈华, 耿直	12	SCI-E	是
合 计							110		

主要完成人（完成单位）：

1. 耿直（北京大学，数学科学学院，统计科学中心）

对本项目重要科学发现的贡献：

成果一：创新性地提出了替代指标悖论，指出现在的确定替代指标准则都不能避免替代指标悖论，论证了避免替代指标悖论的新准则。

成果二：关于 Yule-Simpson 悖论揭示的虚假相关性，研究因果作用的可识别性、混杂因素的形式化定义、判断混杂因素的准则。与完成人 2 联合完成的。

成果三：提出了不可忽略缺失数据机制和不依从性问题的因果推断方法。与完成人 3 联合完成的。

成果四：提出了将大规模的因果网络的分解为小网络结构学习的算法。

2. 郭建华（东北师范大学，数学与统计学院，应用统计教育部重点实验室）

对本项目重要科学发现的贡献：

成果一：关于 Yule-Simpson 悖论揭示的虚假相关性，研究因果作用的可识别性、混杂因素的形式化定义、判断混杂因素的准则。与完成人 1 联合完成的。

成果二：提出了大规模的因果网络的可分解性理论和分解条件。

3. 苗旺（北京大学，光华管理学院，统计科学中心）

对本项目重要科学发现的贡献：

成果一：提出了不可忽略缺失数据机制的可识别性问题和推断方法。与完成人 1 联合完成的。