Engineering 6 (2020) 253–263

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

Causal Inference[†]

Kun Kuang [a,*], Lian Li [b], Zhi Geng [c], Lei Xu [d], Kun Zhang [e], Beishui Liao [f], Huaxin Huang [f], Peng Ding [g], Wang Miao [h], Zhichao Jiang [i]

[a] College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
[b] Department of Computer Science and Technology, Hefei University of Technology, Hefei 230009, China
[c] School of Mathematical Science, Peking University, Beijing 100871, China
[d] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[e] Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[f] School of Humanities, Zhejiang University, Hangzhou 310058, China
[g] University of California Berkeley, Berkeley, CA 94720, USA
[h] Guanghua School of Management, Peking University, Beijing 100871, China
[i] Department of Government & Department of Statistics, Harvard University, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 8 May 2019
Revised 31 July 2019
Accepted 26 August 2019
Available online 8 January 2020

Keywords:
Causal inference
Instructive variables
Negative control
Causal reasoning and explanation
Causal discovery
Counterfactual inference
Treatment effect estimation

ABSTRACT

Causal inference is a powerful modeling tool for explanatory analysis, which might enable current machine learning to become explainable. How to marry causal inference with machine learning to develop explainable artificial intelligence (XAI) algorithms is one of key steps toward to the artificial intelligence 2.0. With the aim of bringing knowledge of causal inference to scholars of machine learning and artificial intelligence, we invited researchers working on causal inference to write this survey from different aspects of causal inference. This survey includes the following sections: "Estimating average treatment effect: A brief review and beyond" from Dr. Kun Kuang, "Attribution problems in counterfactual inference" from Prof. Lian Li, "The Yule–Simpson paradox and the surrogate paradox" from Prof. Zhi Geng, "Causal potential theory" from Prof. Lei Xu, "Discovering causal information from observational data" from Prof. Kun Zhang, "Formal argumentation in causal reasoning and explanation" from Profs. Beishui Liao and Huaxin Huang, "Causal inference with complex experiments" from Prof. Peng Ding, "Instrumental variables and negative controls for observational studies" from Prof. Wang Miao, and "Causal inference with interference" from Dr. Zhichao Jiang.

1. Estimating average treatment effect: A brief review and beyond

Machine learning methods have demonstrated great success in many fields, but most lack interpretability. Causal inference is a powerful modeling tool for explanatory analysis, which might enable current machine learning to make explainable prediction. In this article, we review two classical estimators for estimating causal effect, and discuss the remaining challenges in practice. Moreover, we present a possible way to develop explainable artificial intelligence (XAI) algorithms by marrying causal inference with machine learning.

1.1. The setup

We are interested in estimating the causal effect of a binary variable based on potential outcome framework [1]. For each unit indexed by i = 1, 2, ..., n (n denotes the sample size), we observe a treatment T_i, an outcome, and a vector of observed variables $\boldsymbol{X} \in R^{p\times 1}$, where p refers to the dimension of observed variables. The pair of potential outcomes for each unit i is $\{Y_i(1), Y_i(0)\}$ corresponding to its treatment assignment $T_i = 1$ (treated) or $T_i = 0$ (control). The observed outcome Y_i^{obs} is

$$Y_i^{\text{obs}} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0) \tag{1}$$

Then, the average treatment effect is defined as follows:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

[†] The authors contributed equally to this work. The symbol definitions and notations of each section are relatively independent.

[*] Corresponding author.
E-mail address: kunkuang@zju.edu.cn (K. Kuang).

https://doi.org/10.1016/j.eng.2019.08.016

where function $\mathbb{E}(\cdot)$ denotes the expectation function, and the average treatment effect for treated is defined as $\tau_{\mathrm{t}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$.

To identify τ and τ_{t}, we assume the un-confoundedness—that $T_i \perp [(1),\ Y_i(0)]|\boldsymbol{X}_i$—and assume the overlap of the covariate distribution—that $0 < p(T_i = 1|\boldsymbol{X}_i) < 1$.

1.2. Two estimators

Here, we briefly introduce two of the most promising estimators for treatment effect estimation and discuss them for the case with many observed variables.

1.2.1. Inverse propensity weighting

In fully random experiments, the treatment is randomly assigned to units, implying that $T_i \perp \boldsymbol{X}_i$. In observational studies, however, the treatment T_i is assigned based on $\boldsymbol{X}_i$. To remove the confounding effect from $\boldsymbol{X}_i$, the propensity score, denoted as $e(\boldsymbol{X}_i) = (T_i = 1|\boldsymbol{X}_i)$, was proposed to reweight each unit i. Then, τ can be estimated by the following:

$$\tau = \mathbb{E}\left[\frac{Y_i^{\mathrm{obs}} T_i}{e(\boldsymbol{X}_i)} - \frac{Y_i^{\mathrm{obs}}(1 - T_i)}{1 - e(\boldsymbol{X}_i)}\right] \tag{3}$$

By combining propensity weighting and regression, it is also possible to estimate the treatment effect with a doubly robust method [2]. In high-dimensional settings, not all observed variables are confounders. To address this issue, Kuang et al. [3] suggest separating all observed variables into two parts: the confounders for propensity score estimation, and the adjustment variables for reducing the variance of the estimated causal effect.

1.2.2. Confounder balancing

The other promising way to remove the confounding effect is to balance the distribution of confounders between treated and control groups by sample reweighting with sample weights W, and to estimate τ_{t} as follows:

$$\tau_{\mathrm{t}} = \mathbb{E}\left[Y_i^{\mathrm{obs}}|T_i = 1\right] - \mathbb{E}\left[W_j Y_i^{\mathrm{obs}}|T_j = 0\right] \tag{4}$$

where the sample weights W can be learned by confounder balancing [4] as follows:

$$W = \arg\min_{W} \left\|\mathbb{E}\left[Y_i^{\mathrm{obs}}|T_i = 1\right] - \mathbb{E}\left[W_j Y_i^{\mathrm{obs}}|T_j = 0\right]\right\|_{\infty}^{2} \tag{5}$$

In high-dimensional settings, different confounders can contribute to different confounding biases. Thus, Kuang et al. [5] suggest jointly learning confounder weights for confounders differentiation, learning sample weights for confounder balancing, and simultaneously estimating the treatment effect with a Differentiated Confounder Balancing (DCB) algorithm.

1.3. Remaining challenges

There are now more promising methods available for estimating treatment effect in observational studies, but many challenges remain in making these methods become useful in practice. Here are some of the remaining challenges:

1.3.1. From binary to continuous

The leading estimators are designed for estimating the treatment effect of a binary variable and achieve good performance in practice. In many real applications, however, we care not only about the cause effect of a treatment, but also about the dose response functions, where the treatment dose may take on a continuum of values.

1.3.2. Interaction of treatments

In practice, the treatment can consist of multiple variables and their interactions. In social marketing, the combined causal effects of different advertising strategies may be of interest. More work is needed on the causal analyses of treatment combination.

1.3.3. Unobserved confounders

The existence of unobserved confounders is equivalent to violation of the unconfoundedness assumption and is not testable. Controlling high-dimensional variables may make unconfoundedness more plausible but poses new challenges to propensity score estimation and confounder balancing.

1.3.4. Limited on overlap

Although the overlap assumption is testable, it raises several issues in practice, including how to detect a lack of overlap in the covariate distributions, and how to deal with such a lack, especially in high-dimensional settings. Moreover, estimating the treatment effect is only possible for the region of overlap.

Recently, related works have been proposed to address the above challenges, including continuous treatment [6], the interaction of treatments [7], unobserved confounders [8], and the limits on overlap [9,10].

1.4. Toward causal and stable prediction

The lack of interpretability of most predictive algorithms makes them less attractive in many real applications, especially those requiring decision-making. Moreover, most current machine learning algorithms are correlation based, leading to instability of their performance across testing data, whose distribution might be different from that of the training data. Therefore, it can be useful to develop predictive algorithms that are interpretable for users and stable to the distribution shift from unknown testing data.

By assuming that the causal knowledge is invariant across datasets, a reasonable way to solve this problem is to explore causal knowledge for causal and stable prediction. Inspired by the confounder-balancing techniques from the literature of causal inference, Kuang et al. [11] propose a possible solution for causal and stable prediction. They propose a global variable balancing regularizer to isolate the effect of each individual variable, thus recovering the causation between each variable and response variable for stable prediction across unknown datasets.

Overall, how to deeply marry causal inference with machine learning to develop XAI algorithms is one of key steps toward to the artificial intelligence (AI) 2.0 [12,13], and remains many special issues, challenges and opportunities.

2. Attribution problems in counterfactual inference

In this section, the input variable X and the outcome variable Y are both binary.

Counterfactual inference is an important part of causal inference. Briefly speaking, counterfactual inference is to determine the probability that the event y would not have occurred (y = 0) had the event x not occurred (x = 0), given the fact that event x did occur (x = 1) and event y did happen (y = 1), which can be represented as the following equation:

$$P(y_{x=0} = 0|x = 1,\ y = 1) \tag{6}$$

where $y_{x=0}$ is a counterfactual notion, which denotes the value of y when the setting is x = 0 and the fixing effects of other variables are unchanged, so it is different from the conditional probability $P(y|x = 0)$. This formula reflects the probability that event y will not occur if event x does not occur; that is, it reflects the necessity of the causality of x and y. In social science or logical science, this is called the attribution problem. It is also known as the "but-for" criterion in jurisprudence. The attribution problem has a long history of being studied; however, previous methods used to address this problem have mostly been case studies, statistical analysis, experimental design, and so forth; one example is the influential INUS theory put forward by the Australian philosopher Mackie in the 1960s [14]. These methods are basically qualitative, relying on experience and intuition. With the emergence of big data, however, data-driven quantitative study has been developed for the attribution problem, making the inference process more scientific and reasonable.

Attribution has a twin problem, which is to determine the probability that the event y would have occurred ($y = 1$) had the event x occurred ($x = 1$), given that event x did not occur ($x = 0$) and event y did not happen ($y = 0$). Eq. (7) represents this probability.

$$P(y_{x=1} = 1|x = 0,\ y = 0) \tag{7}$$

This equation reflects the probability that event x causes event y; that is, it reflects the sufficiency of the causality of x and y.

Counterfactual inference corresponds to human introspection, which is a key feature of human intelligence. Inference allows people to predict the outcome of performing a certain action, while introspection allows people to rethink how they could have improved the outcome, given the known effect of the action. Although introspection cannot change the existing *de facto* situation, it can be used to correct future actions. Introspection is a mathematical model that uses past knowledge to guide future action. Unless it possesses the ability of introspection, intelligence cannot be called true intelligence.

Introspection is also important in daily life. For example, suppose Ms. Jones and Mrs. Smith both had cancer surgery. Ms. Jones also had irradiation. Eventually, both recovered. Then Ms. Jones rethought whether she would have recovered had she not taken the irradiation. Obviously, we cannot infer that Ms. Jones would have recovered had she not take the irradiation, based on the fact that Mrs. Smith recovered without irradiation.

There is an enormous amount of this kind of problem in medical disputes, court trials, and so forth. What we are concerned with is what the real causality is, once a fact has occurred for a specific individual case. In these situations, general statistics data—such as the recovery rate with irradiation—cannot provide the explanation. Calculating the necessity of causality by means of introspection and attribution inference plays a key role in these areas [14].

As yet, no general calculation method exists for Eq. (6). In cases that involve solving a practical problem, researchers introduce a monotonic assumption that can be satisfied in most cases; that is:

$$y_{x=1} \geq y_{x=0}$$

The intuition of monotonicity is that the effect y of taking an action ($x = 1$) will not be worse than that of not taking the action ($x = 0$). For example, in epidemiology, the intuition of monotonicity is not true for people who are contrarily infected ($y = 0$) after being quarantined ($x = 1$), and who were uninfected ($y = 1$) before being quarantined ($x = 0$). Because of the monotonicity, Eq. (6) can be rewritten as follows:

$$\begin{aligned} P(y_{x=0} = 0|x = 1,\ y = 1) &= \frac{P(y = 1) - P(y_{x=0} = 1)}{P(x = 1,\ y = 1)} \\ &= \frac{P(y = 1|x - 1) - P(y = 1|x = 0)}{P(y = 1|x - 1)} \\ &\quad + \frac{P(y = 1|x = 0) - P(y_{x=0} = 1)}{P(y = 1|x = 1)} \end{aligned} \tag{8}$$

Eq. (8) has two terms. The first term is named the attributable risk fraction, or the excess risk ratio, and is well known in risk statistics. This term reflects the different risk ratio conditioning on $x = 1$ and $x = 0$. The second term is the confounding factor, which should be particularly noticed. This term reflects the effect confounded by other variables. In a natural environment, a change in y could be caused by x in two different ways: First, it could be directly caused by a change in x; or, second, it could be caused by other variables. This phenomenon is called confounding. The difference $P(y = 1|x = 0) - P(y_{x=0} = 1)$ denotes the degree of confounding. In some situations, the change in x did give rise to the change in y, but x may not be the reason for the change in y (e.g., the sun rises after the cock crows). It is possible to exclude confounding by means of scientific experiments to determine the true causality of the change in y. However, scientific experiments can hardly be conducted in many social science problems, or even in some natural science problems. In such cases, only the observational data can be obtained. Thus, the question of how to recognize confounding from observational data in order to determine the true causality is a fundamental problem in artificial intelligence.

In order to explain the relationship between the attributable risk fraction and the confounding factor, and their roles in the attribution problem (i.e., the necessity of causality) more specifically, we applied the example in Ref. [15]. In this example, Mr. A goes to buy a drug to relieve his pain and dies after taking the drug. The plaintiff files a lawsuit to ask the manufacturer to take responsibility. The manufacturer and plaintiff provide the drug test results (i.e., experimental data) and survey results (i.e., nonexperimental data), respectively. The data is illustrated in Table 1, where $x = 1$ denotes taking drugs, while $y = 1$ denotes death.

The manufacturer's data comes from strict drug safety experiments, while the plaintiff's data comes from surveys among patients taking drugs by their own volition. The manufacturer claims that the drug was approved based on the drug distribution regulations. Although it causes a minor increase in death rate (from 0.014 to 0.016), this increase is acceptable compared with the analgesic effect. Based on the traditional calculation of the attributable risk fraction (excess risk ratio), the responsibility taken by the manufacturer is

$$\frac{P(y = 1|x = 1) - P(y = 1|x = 0)}{P(y = 1|x = 1)} = \frac{0.016 - 0.014}{0.016} = 0.125 \tag{9}$$

The plaintiff argues that the drug test was conducted under experimental protocols, the subjects were chosen randomly, and the subjects did not take the drug of their own volition. Therefore, there is bias in the experiment, and the experimental setting differs from the actual situation. There is a huge difference between observational data and experimental data. Given the fact of the death of Mr. A, the calculation of the manufacturer's responsibility should obey the counterfactual equation. The result is

$$\begin{aligned} &\frac{P(y=1|x-1) - P(y=1|x=0)}{P(y=1|x-1)} + \frac{P(y=1|x=0) - P(y_{x=0}=1)}{P(y=1|x=1)} \\ &= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = 1 \end{aligned} \tag{10}$$

Therefore, the manufacturer should take full responsibility for the death of Mr. A.

Table 1
Experimental and non-experimental data for the example of a drug lawsuit.

Outcomes	Experimental data (number of patients)		Non-experimental data (number of patients)	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$
Deaths ($y = 1$)	16	14	2	28
Survivals ($y = 0$)	984	986	998	972

A quick look shows that, based on the survey data, the death rates of taking and not taking the drug are 0.2% and 2.8%, respectively, which is in favor of the manufacturer. However, after careful analysis, the confounding factor is $P(y=1|x=0)-P(y_{x=0}=1)=0.014$; that is, half of the subjects died due to reasons other than not taking the drug. This part should not be attributed to the drug, so the manufacturer's responsibility increases. Of course, there is some doubt regarding whether the manufacturer should take full responsibility, as well as regarding the rationality and scientificity of the calculation [16]. Nevertheless, this example demonstrates that there are confounding factors that will disturb the discovery of true causality. The question of how to determine confounding factors is a practical problem in causal inference, naturally, and is also important in counterfactual inference.

In data science, there are simulated data and objective data, with the latter containing experimental data and observational data. Although observational data are objective, easily available, and low in cost, the confounding problems among them become an obstacle for causal inference [17]. In particular, there may be unknown variables (i.e., hidden variables) in an objective world. These variables are not observed, but may have effects on known variables—that is, the known variables should be sensitive to unmeasured confounding due to unknown variables. In this aspect, current studies on confounding are still in their infancy. Readers can refer to Ref. [18] for more detail.

3. The Yule–Simpson paradox and the surrogate paradox

An association measurement between two variables may be dramatically changed from positive to negative by omitting a third variable, Z; this is called the Yule–Simpson paradox [19,20]. The third variable, Z, is called a confounder. A numerical example is shown in Table 2. The risk difference (RD) is the difference between the proportion of lung cancer in the smoking group and that in the no-smoking group, RD = (80/200) − (100/200) = −0.10, which is negative. If the 400 persons listed in Table 2 are split into males and females, however, a dramatic change can be seen (Table 3). The RDs for both males and females are positive, at 0.10. This means that while smoking is bad for both males and females, separately, smoking is good for all of these persons.

The main difference between causal inference and other forms of statistical inference is whether the confounding bias induced by the confounder is considered. For experimental studies, it is possible to determine which variables affect the treatment or exposure; this is particularly true for a randomized experiment, in which the treatment or exposure is randomly assigned to individuals, as there is no confounder affecting the treatment. Thus, randomized experiments are the gold standard for causal inference. For observational studies, it is key to observe a sufficient set of confounders or an instrumental variable that is independent of all confounders. However, neither a sufficient confounder set nor an instrumental variable can be verified by observational data without manipulations.

In scientific studies, a surrogate variable (e.g., a biomarker) is often measured instead of an endpoint, due to its infeasible measurement; and, then, the causal effect of a treatment on the unmeasured endpoint is predicted by the effect on the surrogate. The surrogate paradox means that the treatment has a positive effect on the surrogate, and the surrogate has a positive effect on the endpoint, but the treatment may have a negative effect on the endpoint [21]. Numerical examples are given in Refs. [21,22]. This paradox also queries whether scientific knowledge is useful for policy analysis [23]. As a real example, doctors have the knowledge that an irregular heartbeat is a risk factor for sudden death. Several therapies can correct irregular heartbeats, but they increase mortality [24].

Table 2
Smoking and lung cancer.

Condition	Number of persons		
	Cancer	No cancer	Total
Smoking	80	120	200
No smoking	100	100	200

Table 3
Smoking and lung cancer with populations stratified by gender.

Condition	Males		Females	
	Cancer	No cancer	Cancer	No cancer
Smoking	35	15	45	105
No smoking	90	60	10	40

Yule–Simpson paradox and the surrogate paradox warn about that a conclusion obtained from data can be inverted due to unobserved confounders and emphasize the importance of using appropriate approaches to obtain data. To avoid the Yule–Simpson paradox, first, randomization is the golden standard approach for causal inference. Second, the use of an experimental approach to obtain data is expected, if randomization is prohibited, as such an approach attempts to balance all possible unobserved confounders between the two groups to be compared. Third, an encouragement-based experimental approach—in which benefits are randomly assigned to a portion of the involved persons, such that the assignment can change the probability of their exposure—can be used to design an instrumental variable. Finally, for a pure observational approach, it is necessary to verify the assumptions required for causal inference using field knowledge, and to further execute a sensitivity analysis for violations of these assumptions. The two paradoxes also point out that a syllogism and transitive reasoning may not be applicable to statistical results. Statistically speaking, smoking is good for both males and females, and the studied population consists of these males and females; however, the statistics indicate that smoking is bad for the population as a whole. Statistics may show that a new drug can correct irregular heartbeats, and it is known that a regular heartbeat can promote survival time, both statistically speaking and for individuals; however, the new drug may still shorten the survival time of these persons in terms of statistics.

4. Causal potential theory

Extensive efforts have been made to detect causal direction, evaluate causal strength, and discover causal structure from observations. Examples include not only the studies based on conditional independence and directed acyclic graphs (DAGs) by Pearl, Spirtes, and many others, but also those on the Rubin causal model (RCM), structural equation model (SEM), functional causal model (FCM), additive noise model (ANM), linear non-Gaussian acyclic model (LiNGAM), post-nonlinear (PNL) model, and causal generative neural networks (CGNNs), as well as the studies that discovered star structure [25] and identified the so called ρ-diagram [26]. To some extent, these efforts share a similar direction of thinking. First, one presumes a causal structure (e.g., merely one direction in the simplest case, or a DAG in a sophisticated situation) for a multivariate distribution, either modeled in parametric form or partly inspected via statistics, which is subject to certain constraints. Second, one uses observational data to learn the parametric model or estimate the statistics, and then examines whether the model fits the observations and the constraints are satisfied; based on this, one verifies whether the presumed causal structure externally describes observations the well. Typically, a set of causal structures are presumed as candidates, among which the best is selected.

Causal potential theory (CPT) was recently proposed as a very different way of thinking [27]. In analogy to physics, causality is here regarded as an intrinsic kinetic nature caused by a causal potential energy. Without losing generality, this CPT is introduced by starting with the consideration of a cause-effect relation between a pair of variables, x, y,[†] in an environment, U. Instead of presuming a causal structure (i.e., a specific direction), one estimates a nonparametric distribution $p_U(x,\ y) \triangleq p(x,\ y|U)$ from samples of x, y, and obtains the corresponding causal potential energy $E_U(x,\ y) \propto -\ln p_U(x,\ y)$ in an analogy based on the Gibbs distribution. In such a perspective of causal dynamics, an event occurring at x, y is associated with $E_U(x,\ y)$ that yields a force $[g_x,\ g_y]$ to cause subsequent events by the dynamics $[\dot{x}_t,\ \dot{y}_t] \propto -[g_x,\ g_y]$, driving the information flow or causal process toward an area with the lowest energy or, equivalently, toward an area in which events have high chances to occur, using the notations $g_U \triangleq \nabla_U E_U$ and $\dot{u}_t \triangleq \mathrm{d}u/\mathrm{d}t$. That is, CPT regards causality as an intrinsic nature of the dynamics $[\dot{x}_t,\ \dot{y}_t] \propto -[g_x,\ g_y]$ and discovers causality by analyzing $[g_x,\ g_y]$.

Table 4 shows two roads for analyzing CPT causality. Road$_\mathrm{A}$ is proceeded by testing a "Yes" or "No" answer on the mutual independence between g_y, y and on that between g_x, x, resulting in four types of Y-N combinations. The first two types indicate two types of causality. The third type, Y-Y, indicates the independence between x, y—that is, indicates that there is no relation between them. The last type, N-N, indicates "unclear ?"—that is, further study is needed to determine whether a causal relation still occurs locally, or even reciprocally, in some regions of x, y, although there is no causal relation detected globally between x, y. Road$_\mathrm{A}$ needs an independence test. In contrast, Road$_\mathrm{B}$ turns the problem into supervised learning, with x, y as inputs into a neural net to fit two gradient components $[g_x, g_y]$, each of which is fit by a different neural net, with one or both of x, y as inputs, respectively. An appropriate one is chosen according to not only fit, but also simplicity. Table 4 lists four types of outcomes based on this method [27].

It is possible to seek a certain estimator to obtain g_x, g_y directly from samples x_t, y_t, where $t = 1, \ldots, N$ and N refers to the sample size. It is also possible to obtain g_x, g_y indirectly, by estimating $p_U(x,\ y)$ first; that is, by performing a kernel estimate $p_h(x,\ y) = \frac{1}{N}\sum_{t=1}^{N} G\left(x,\ y\middle|x_t,\ y_t,\ h^2\boldsymbol{I}\right)$, where there is a Gaussian of mean m and variance σ^2. Alternatively, it is possible to obtain p_U by one presumed causal structure, and to perform CPT analyses on this p_U.

Experiments on the CauseEffectPairs (CEP) benchmark have demonstrated that a preliminary and simple implementation of CPT has achieved performances that are comparable with ones achieved by state-of-art methods.

Further development is to explore the estimation of causal structure between multiple variable distributions and multiple variables, possibly along two directions. One is simply integrating the methods in Table 4 into the famous Peter–Clark (PC) algorithm [28], especially on edges that are difficultly identified by independent and conditional independent tests. The other is turning the conditions that g_y is uncorrelated (or independent) of x and that g_x is uncorrelated (or independent) of y into multivariate polynomial equations, and adding the equations into the ρ-diagram equations in Ref. [26], e.g., Eq. (29) and Eq. (33), to get an augmented group of polynomial equations. Then, the well known Wen-Tsun Wu method may be adopted to check whether the equations have unique or a finite number of solutions.

5. Discovering causal information from observational data

Causality is a fundamental notion in science, and plays an important role in explanation, prediction, decision-making, and control [28,29]. There are two essential problems to address in modern causality research. One essential problem is the identification of causal effects, that is, identifying the effects of interventions, given the partially or completely known causal structure and some observed data; this is typically known as "causal inference." For advances in this research direction, readers are referred to Ref. [29] and the references therein. In causal inference, causal structure is assumed to be given in advance—but how can we find causal structure if it is not given? A traditional way to discover causal relations resorts to interventions or randomized experiments, which are too expensive or time-consuming in many cases, or may even be impossible from a practical standpoint. Therefore, the other essential causality problem, which is how to reveal causal information by analyzing purely observational data, has drawn a great deal of attention [28].

In the last three decades, there has been a rapid spread of interest in principled methods causal discovery, which has been driven in part by technological developments. These technological developments include the ability to collect and store big data with huge numbers of variables and sample sizes, and increases in the speed of computers. In domains containing measurements such as satellite images of weather, functional magnetic resonance imaging (fMRI) for brain imaging, gene-expression data, or single-nucleotide polymorphism (SNP) data, the number of variables can range in the millions, and there is often very limited background knowledge to reduce the space of alternative causal hypotheses. Causal discovery techniques without the aid of an automated search then appear to be hopeless. At the same time, the availability of faster computers with larger memories and disc space allow for practical implementations of computationally intensive automated algorithms to handle large-scale problems.

It is well known in statistics that "causation implies correlation, but correlation does not imply causation." Perhaps it is fairer to say that correlation does not *directly* imply causation; in fact, it has become clear that under suitable sets of assumptions, the causal structure (often represented by a directed graph) underlying a set of random variables can be recovered from the variables' observed data, at least to some extent. Since the 1990s, conditional

Table 4
Two roads for analyzing CPT causality.

$\nabla_U E_U$	$y \to x$		$x \to y$		$x \perp\!\!\!\perp y$		$x?y$	
	Road$_\mathrm{A}$	Road$_\mathrm{B}$	Road$_\mathrm{A}$	Road$_\mathrm{B}$	Road$_\mathrm{A}$	Road$_\mathrm{B}$	Road$_\mathrm{A}$	Road$_\mathrm{B}$
g_x	Dependent of y	$\xi(x, y) + \varepsilon$	$\perp\!\!\!\perp y$	$\xi(x) + \varepsilon$	$\perp\!\!\!\perp y$	$\xi(x) + \varepsilon$	Dependent of y	$\xi(x,y) + \varepsilon$
g_y	$\perp\!\!\!\perp x$	$\eta(y) + \varepsilon$	Dependent of x	$\eta(x,y) + \varepsilon$	$\perp\!\!\!\perp x$	$\eta(y) + \varepsilon$	Dependent of x	$\eta(x,y) + \varepsilon$

[†] In this section, we reuse x, y to denote a pair variable, their relationship might be cause and effect.

independence relationships in the data have been used for the purpose of estimating the underlying causal structure. Typical (conditional independence) constraint-based methods include the PC algorithm and fast causal inference (FCI) [28]. Under the assumption that there is no confounder (i.e., unobserved direct common cause of two measured variables), the result of PC is asymptotically correct. FCI gives asymptotically correct results even when there are confounders. These methods are widely applicable because they can handle various types of causal relations and data distributions, given reliable conditional independence testing methods. However, they may not provide all the desired causal information, because they output (independence) equivalence classes—that is, a set of causal structures with the same conditional independence relations. The PC and FCI algorithms output graphical representations of the equivalence classes. In cases without confounders, there also exist score-based algorithms that estimate causal structure by optimizing some properly defined score function. The greedy equivalence search (GES), among them, is a widely used two-phase procedure that directly searches over the space of equivalence classes.

In the past 13 years, it has been further shown that algorithms based on properly constrained FCMs are able to distinguish between different causal structures in the same equivalence class, thanks to additional assumptions on the causal mechanism. An FCM represents the outcome or effect variable Y as a function of its direct causes X and some noise term E, that is, $Y = f(X, E)$, where E is independent of X. It has been shown that, without constraints on function f, for any two variables, one of them can always be expressed as a function of the other and independent noise [30]. However, if the functional classes are properly constrained, it is possible to identify the causal direction between X and Y because for wrong directions, the estimated noise and hypothetical cause cannot be independent (although they are independent for the right direction). Such FCMs include the LiNGAM [31], where causal relations are linear and noise terms are assumed to be non-Gaussian; the post-nonlinear (PNL) causal model [32], which considers nonlinear effects of causes and possible nonlinear sensor/measurement distortion in the data; and the nonlinear ANM [33,34], in which causes have nonlinear effects and noise is additive. For a review of these models and corresponding causal discovery methods, readers are referred to Ref. [30].

Causal discovery exploits observational data. The data are produced not only by the underlying causal process, but also by the sampling process. In practice, for reliable causal discovery, it is necessary to consider specific challenges posed in the causal and sampling processes, depending on the application domain. For example, for multivariate time series data such as mRNA expression series in genomics and blood-oxygenation-level-dependent (BOLD) time series in neuropsychology, finding the causal dynamics generating such data is challenging for many reasons, including nonlinear causal interactions, a much lower data-acquisition rate compared with the underlying rates of change, feedback loops in the causal model, the existence of measurement error, nonstationarity of the process, and possible unmeasured confounding causes. In clinical studies, there is often a large amount of missing data. Data collected on the Internet or in hospital often suffer from selection bias. Some datasets involve both mixed categorical and continuous variables, which may pose difficulties in conditional independence tests and in the specification of appropriate forms of the FCM. Many of these issues have recently been considered, and corresponding methods have been proposed to address them.

Causal discovery has benefited a great deal from advances in machine learning, which provide an essential tool to extract information from data. On the other hand, causal information describes properties of the process that render a set of constraints on the data distribution and is able to facilitate understanding and solve a number of learning problems involving distribution shift or concerning the relationship between different factors of the joint distribution. In particular, for learning under data heterogeneity, it is naturally helpful to learn and model the properties of data heterogeneity, which then benefit from causal modeling. Such learning problems include domain adaptation (or transfer learning) [35], semi-supervised learning, and learning with positive and unlabeled examples. Leveraging causal modeling for recommender systems and reinforcement learning is becoming an active research field in recent years.

6. Formal argumentation in causal reasoning and explanation

In this section, we sketch why and how formal argumentation can play an important role in causal reasoning and explanation. Reasoning in argumentation is realized by constructing, comparing, and evaluating arguments [36]. An argument commonly consists of a claim that may be supported by premises, which can be observations, assumptions, or intermediate conclusions of some other arguments. The claim, the premises, and the inference relation between them may be the subject of rebuttals or counterarguments [37]. An argument can be accepted only when it survives all attacks. In AI, formal argumentation is a general formalism for modeling defeasible reasoning. It provides a natural way for justifying and explaining causation, and is complementary to machine learning approaches, for learning, reasoning, and explaining cause-and-effect relations.

6.1. Nonmonotonicity and defeasibility

Causal reasoning is the process of identifying causality, that is, the relationship between a cause and its effect, which is often defeasible and nonmonotonic. On the one hand, causal rules are typically defeasible. A causal rule may be represented in the form “c causes e” where e is some effect and c is a possible cause. The causal connective is not a material implication, but a defeasible conditional with strength or uncertainty. For example, “turning the ignition key causes the motor to start, but it does not imply it, since there are some other factors such as there being a battery, the battery not being dead, there being gas, and so on” [38]. On the other hand, causal reasoning is nonmonotonic, in the sense that causal connections can be drawn tentatively and retracted in light of further information. It is usually the case that c causes e, but c and d jointly do not cause e. For example, an agent believes that turning the ignition key causes the motor to start, but when it knows that the battery is dead, it does not believe that turning the ignition key will cause the motor to start. In AI, this is the famous qualification problem. Since the potentially relevant factors are typically uncertain, it is not cost effective to reason explicitly. So, when doing causal inference, people usually “jump” to conclusions and retract some conclusions when needed. Similarly, reasoning from evidence to cause is nonmonotonic. If an agent observes some effect e, it is allowed to hypothesize a possible cause c. The reasoning from the evidence to a cause is abductive, since for some evidence, one may accept an abductive explanation if no better explanation is available. However, when new explanations are generated, the old explanation might be discarded.

6.2. Efficiency and explainability

From a perspective of computation, monotonicity is a crucial property of classical logic, which means that each conclusion obtained by local computation using a subset of knowledge is equal to the one made by global computation using all the knowledge. This property does not hold in nonmonotonic reasoning and, therefore, the computation could be highly inefficient. Due to the nonmonotonicity of causal reasoning, in order to improve efficiency, formal argumentation has been evidenced to be a good candidate, by comparing it with some other nonmonotonic formalisms such as default logic and circumscription. The reason is that in formal argumentation, computational approaches may take advantage of the divide-and-conquer strategy and maximal usage of existing computational results in terms of the reachability between nodes in an argumentation graph [39]. Another important property of causal reasoning in AI is explainability. Traditional nonmonotonic formalisms are not ideal for explanation, since all the proofs are not represented in a human understandable way. Since the purpose of explanation is to let the audience understand, the cognitive process of comparing and contrasting arguments is significant [37]. Argumentation provides such a way by exchanging arguments in terms of justification and argument dialogue [40].

6.3. Connections to machine learning approaches

In explainable AI, there are two components: the explainable model and the explanation interface. The latter includes reflexive explanations that arise directly from the model and rational explanations that come from reasoning about the user's beliefs. To realize this vision, it is natural to combine argumentation and machine learning, in the sense that knowledge is obtained by machine learning approaches, while the reasoning and explanation are realized by argumentation. Since argumentation provides a general approach for various kinds of reasoning in the context of disagreement, and can be combined with some uncertainty measures, such as probability and fuzziness, it is very flexible to model the knowledge learned from data. An example is when a machine learns features and produces an explanation, such as "This face is angry, because it is similar to these examples, and dissimilar from those examples." This is an argument, which might be attacked by other arguments. And, in order to measure the uncertainty described by some words such as "angry," one may choose to use possibilistic or probabilistic argumentation [41]. Different explanations may be in conflict. For instance, there could be some cases invoking specific examples or stories that support a choice, and rejections of an alternative choice that argue against less-preferred answers based on analytics, cases, and data. By using argumentation graphs, these kinds of support-and-attack relations can be conveniently modeled and can be used to compute the status of conflicting arguments for different choices.

7. Causal inference with complex experiments

The potential outcomes framework for causal inference starts with a hypothetical experiment in which the experimenter can assign every unit to several treatment levels. Every unit has potential outcomes corresponding to these treatment levels. Causal effects are comparisons of the potential outcomes among the same set of units. This is sometimes called the experimentalist's approach to causal inference [42]. Readers are referred to Refs. [43–46], for textbook discussions.

7.1. Randomized factorial experiments

Splawa-Neyman [47] first formally discussed the following randomization model. In an experiment with n units, the experimenter randomly assigns $(n_1, \ldots, n_J)$ units to treatment levels $(1, \ldots, J)$, where $n = \sum_{j=1}^{J} n_j$. Unit i has potential outcomes $\{Y_i(1), \ldots, Y_i(J)\}$, with $Y_i(j)$ being the hypothetical outcome if unit i receives treatment level j. With potential outcomes, we can define causal effects; for example, the comparison between treatment levels j and j' as $\tau(j, j') = n^{-1}\sum_{i=1}^{n}\{Y_i(j) - Y_i(j')\}$. Let $T_i(j)$ be the indicator if unit i actually receives treatment level j. Let $Y_i = \sum_{j=1}^{J} T_i(j) Y_i(j)$ be the observed outcome of unit i. With observed data $\{T_i(1), \ldots, T_i(J), Y_i\}_{i=1}^{n}$, Splawa-Neyman [47] proposed to use $\widehat{\tau}(j, j') = n_j^{-1}\sum_{i=1}^{n} T_i(j) Y_i - n_{j'}^{-1}\sum_{i=1}^{n} T_i(j') Y_i$ as an estimator for $\tau(j, j')$. He showed that $\widehat{\tau}(j, j')$ is unbiased with variance $\frac{S^2(j)}{n_j} + \frac{S^2(j')}{n_{j'}} - \frac{S^2(j-j')}{n}$, where $S^2(j)$, $S^2(j')$ and $S^2(j-j')$ are the sample variances of $Y_i(j)$, $Y_i(j')$ and $Y_i(j) - Y_i(j')$. Note that the randomness comes from the treatment indicators with all the potential outcomes fixed. Splawa-Neymanhas [47] further discussed variance estimation and the large-sample confidence interval.

We can extend the framework from Ref. [47] to a general causal effect defined as $\tau = n^{-1}\sum_{i=1}^{n}\tau_i$ where $\tau_i = \sum_{j=1}^{J} c_j Y_i(j)$ is the individual effect and the c_j are contrast matrices with $\sum_{j=1}^{J} c_j = 0$. With appropriately chosen contrast matrices, the special cases include analysis of variance [48] and factorial experiments [49,50]. Furthermore, with an appropriately chosen subset of units, the special cases include subgroup analysis, post-stratification [51], and peer effects [52]. Ref. [53] provides the general forms of central limit theorems under this setting for asymptotic inference. Ref. [54] discusses split-plot designs, and Ref. [55] discusses general designs.

7.2. The role of covariates in the analysis of experiments

Splawa-Neyman randomization model [47] also allows for the use of covariates to improve efficiency without strong modeling assumptions. In the case with a binary treatment, for unit i, let $\{Y(1), Y(0)\}$ be the potential outcomes, T_i be the binary treatment indicator, and x_i be pretreatment covariates. The average causal effect $\tau = n^{-1}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}$ has an unbiased estimator $\widehat{\tau} = n_1^{-1}\sum_{i=1}^{n} T_i Y_i - n_0^{-1}\sum_{i=1}^{n}(1 - T_i) Y_i$. Fisher [56] suggested using the analysis of covariance to improve efficiency; that is, running a least squares fit of Y_i on T_i and x_i and using the coefficient of T_i to estimate τ. Ref. [57] uses the model from Ref. [47] to show that Fisher's analysis of the covariance estimator is inferior because it can be even less efficient than $\widehat{\tau}$ and the ordinary least squares can give an inconsistent variance estimate. Ref. [58] proposes a simple correction: First, center covariates to have mean $\bar{x} = 0$; second, run a least squares fit of Y_i on $(T_i, x_i, T_i \times x_i)$ and use the coefficient of T_i to estimate τ, and third, use the Eicker–Huber–White variance estimator [59–61]. With large samples, the estimator from Ref. [58] is at least as efficient as $\widehat{\tau}$, and that researcher's variance estimate is consistent for the true variance of $\widehat{\tau}$.

Ref. [62] extends to the setting with high-dimensional covariates and replaces the least squares fit by the least absolute shrinkage and selection operator (LASSO) [63]. Ref. [64] examines the theoretical boundary of the estimator from Ref. [58], allowing for a diverging number of covariates. Ref. [65] investigates treatment effect heterogeneity using the least squares fit of Y_i on $(T_i, x_i, T_i \times x_i)$. Ref. [66] discusses covariate adjustment in a factorial experiment, and Ref. [67] discusses covariate adjustment in general designs.

7.3. The role of covariates in the design of experiments

An analyzer can use covariates to improve the estimation efficiency. As a dual, a designer can use covariates to improve the covariate balance and consequently improve the estimation efficiency. Ref. [68] hints at the idea of re-randomization—that is, only accepting random allocation that ensures covariate balance. In particular, we accept a random allocation $(T_1, \ldots, T_n)$ if and only if

$\widehat{\tau}'_x\left\{\frac{nS_x^2}{(n_1 n_0)}\right\}^{-1}\widehat{\tau}_x \leq a$, where $\widehat{\tau}_x = n_1^{-1}\sum_{i=1}^n T_i x_i - n_0^{-1}\sum_{i=1}^n (1-T_i)x_i$, $S_x^2 = (n-1)^{-1}\sum\limits_{i=1}^{n}(x_i-\hat{x})(x_i-\hat{x})'$and $a > 0$ is a predetermined constant. Ref. [69] formally discusses its statistical properties under the constant treatment effect model with equal group sizes and Gaussian covariates. Ref. [70] develops its asymptotic theory without these assumptions. In particular, Ref. [70] shows that $\widehat{\tau}$ has a non-Gaussian limiting distribution and is more concentrated at τ under re-randomization than under complete randomization. A consequence of the result from Ref. [70] is that when $a \approx 0$, the asymptotic variance of $\widehat{\tau}$ under re-randomization is identical to the estimator from Ref. [58] under complete randomization. Therefore, we can view re-randomization as the dual of regression adjustment.

Ref. [71] proposes a re-randomization scheme that allows for tiers of covariates, and Ref. [70] derives its asymptotic properties. Refs. [72,73] extend re-randomization to factorial experiments, and Ref. [74] proposes sequential re-randomization.

7.4. Final remarks

Following Ref. [47], I have focused on the repeated sampling properties of estimators with randomized experiments. Alternatively, Fisher randomization tests are finite-sample exact for any test statistics and for any designs, under the sharp null hypothesis that $Y_i(1) = \cdots = Y_i(J)$ for all units i = 1, ..., J [46,75,76]. Refs. [77,78] propose the use of covariate adjustment in randomization tests, and Ref. [69] proposes the use of randomization tests to analyze re-randomization. Refs. [79–81] apply randomization tests to experiments with interference. Refs. [48,50,82] discuss the properties of randomization tests for weak null hypotheses. Refs. [83–85] invert randomization tests to construct exact confidence intervals. Finally, Ref. [86] discusses different inferential frameworks from the missing data perspective.

8. Instrumental variables and negative controls for observational studies

In a great deal of scientific research, the ultimate goal is to evaluate the causal effect of a given treatment or exposure on a given outcome or response variable. Since the work published in Ref. [75], randomized experiments have become a powerful and influential tool for the evaluation of causal effects; however, they are not feasible in many situations due to ethical issues, expensive cost, or imperfect compliance. In contrast, observational studies offer an important source of data for scientific research. However, causal inference with observational studies is challenging, because confounding may arise. Confounders are covariates that affect both the primary exposure and the outcome. In the presence of unmeasured confounders, statistical association does not imply causation, and vice versa, which is known as the Yule–Simpson paradox [19,20]. Refs. [87,88] review the concepts of confounding, and Refs. [2,89,90] discuss methods for the adjustment of observed confounders, such as regressing analysis, propensity score, and inverse probability weighting, as well as doubly robust methods. Here, we review two methods for the adjustment of unmeasured confounding: the instrumental variable approach and the negative control approach.

Throughout, we let X and Y denote the exposure and outcome of interest, respectively, and we let U[†] denote an unmeasured confounder; for simplicity, we omit observed confounders, which can be incorporated in the following by simply conditioning on them. We use lowercase letters to denote realized values of random variables—for example, y for a realized value of Y.

The instrumental variable approach, which was first proposed in econometrics literature in the 1920s [91,92], has become a popular method in observational studies to mitigate the problem of unobserved confounding. In addition to the primary treatment and outcome, this approach involves an instrumental variable Z that satisfies three core assumptions:

(1) It has no direct effect on the outcome, that is, $\mathrm{Z} \perp \mathrm{Y}|(\mathrm{X},\ \mathrm{U})$ *(exclusion restriction);*
(2) It is independent of the unobserved confounder, that is, $\mathrm{Z} \perp \mathrm{U}$ *(independence);*
(3) It is associated with the exposure, that is, $\mathrm{Z} \not\perp \mathrm{X}$ *(relevance).*

Under these three assumptions, only certain upper and lower bounds of causal effects can be derived [93,94], and extra model assumptions are required to achieve identification. The SEM [91,95] and structural mean model [96] are commonly used models, which in fact can achieve identification by assuming effect homogeneity (see Section 16 of Ref. [97]). One such example is the linear regression model $E(Y|X,\ U) = \alpha + \beta X + U$, which encodes a constant causal effect in the regression coefficient β and yields the well-known instrumental variable identification $\beta_{\mathrm{iv}} = \sigma_{zy}/\sigma_{xz}$. Alternatively, in certain situations, especially when Z is a binary treatment assignment that occurs before X, it is sometimes reasonable to assume effect monotonicity: The effect of Z on X is monotone, that is, $X_{Z=1} \geq X_{Z=0}$, which means that no one accepts the opposite treatment of this assignment. The monotonicity assumption leads to identification of the complier average causal effect (CACE) = $E(Y_1 - Y_0 \mid X_1 = 1, X_0 = 0)$, as shown in Ref. [98]. As an extension of the single instrument case, Refs. [99,100] consider variable selection and estimation with high-dimensional instrumental variables.

However, in practice, the instrumental variable assumptions may not be met, and the approach is highly sensitive to the violation of any of them. Validity checking and violation detection of these assumptions are important before applying the instrumental variable approach, and have been attracting researchers' attention [94,101]. In case of a violation of the core assumptions, identification of the causal effect is often impossible, and bounding and sensitivity analysis methods [102,103] have been proposed for causal inference.

Alternatively, we have formally established the double negative control method [104–106] for the adjustment of unmeasured confounding. The negative control approach we have proposed also offers a promising mitigation tool for invalid instrumental variables. Negative control variables are classified into two classes: negative control outcome $W: W \perp X|U,\ W \not\perp U$ and negative control exposure $Z: Z \perp Y|(U,X),\ Z \perp W|(U,X)$. The negative control exposure Z can be viewed as a generalization of an instrumental variable that fails to be independent of the unmeasured confounder, and the negative control outcome W is used to eliminate the bias. Given both a negative control exposure and outcome, Refs. [104,106] show that the average causal effect is nonparametrically identified under certain regularity conditions. For illustration, consider again the regression model $E(Y|X,\ U) = \alpha + \beta X + U$, and assume that $E(W|U)$ also follows a linear model; then, β can be identified by the following:

$$\beta_{\mathrm{nc}} = \frac{\sigma_{xw}\sigma_{zy} - \sigma_{xy}\sigma_{zw}}{\sigma_{xw}\sigma_{xz} - \sigma_{xx}\sigma_{zw}}$$

This formula does apply to a valid instrumental variable; in which case, $Z \perp U$, and thus, σ_{zw} = 0, according to the negative control outcome assumption. Therefore, the instrumental variable identification can be viewed as a special case of the negative con-

[†] In this section, we reuse U to denote the unmeasured confounders. Please note that, U was used to denote an environment in Section 4.

trol approach. However, in contrast to the instrumental variable, negative controls require weak assumptions that are more likely to hold in practice. Refs. [107,108] provide elegant surveys on the existence of negative controls in observational studies. Refs. [105,109] point out that negative controls are widely available in time series studies, as long as no feedback effect is present, such as studies about air pollution and public health.

Refs. [107,109,110] examine the use of negative controls for confounding detection or bias reduction when a solely negative control exposure or outcome is available but are unable to achieve identification. Refs. [111,112] propose the use of multiple negative control outcomes to remove confounding in statistical genetics but must rest on a factor analysis model.

9. Causal inference with interference

The stable unit treatment value assumption plays an important role in the classical potential outcomes framework. It assumes that there is no interference between units [76]. However, interference is likely to be present in many experimental and observational studies, where units socially or physically interact with each other. For example, in educational or social sciences, people enrolled in a tutoring or training program may have an effect on those not enrolled due to the transmission of knowledge [113,114]. In epidemiology, the prevention measures for infectious diseases may benefit unprotected people by reducing the probability of contagion [115,116]. In these studies, one unit's treatment can have a direct effect on its own outcome as well as a spillover effect on the outcome of other units. The direct and spillover effects are of scientific or societal interest in real problems; they enable an understanding of the mechanism of a treatment effect, and provide guidance for policy making and implementation.

In the presence of interference, the number of potential outcomes of a unit grows exponentially with the number of units.[†] As a result, it is intractable to estimate the direct and spillover effects without restriction in the literature on the estimation of treatment effects with interference structure. There has been a rapidly growing interest in interference (see Ref. [117] for a recent review). A significant direction of work focuses on limited interference within nonoverlapping clusters and assumes that there is no interference between clusters [52,114,118–122]. This is referred to as the partial interference assumption [114]. Recently, several researchers have considered the relaxation of the partial interference assumption to account for a more general structure of interference (e.g., Refs. [123–126]). The variance estimation is more complicated under interference. As pointed out in Ref. [118], it is difficult to calculate the variances for the direct and spillover effects even under partial interference. In model-free settings, a typical assumption for obtaining valid variance estimation is that the outcome of a unit depends on the treatments of other units only through a function of the treatments. Ref. [118] provides a variance estimator under the stratified interference assumption, and Ref. [124] generalizes it under a weaker assumption.

Another direction of work targets new designs to estimate treatment effects based on the interference structure. Under the partial interference assumption, Ref. [118] proposes the two-stage randomized experiment as a general experimental solution to the estimation of the direct and spillover effects. In more complex structures such as social networks, researchers have proposed several designs for the point and variance estimation of the treatment effects [127–129].

For the inference under interference, Refs. [130,131] rely on models for the potential outcomes. Ref. [79] develops a conditional randomization test for the null hypothesis of no spillover effect. Ref. [80] extends this test to a larger class of hypotheses restricted to a subset of units, known as focal units. Building on this work, Ref. [132] provides a general procedure for obtaining powerful conditional tests.

Interference brings up new challenges. First, the asymptotic properties require advanced techniques deriving. Ref. [133] investigates the consistency of the difference in the means estimator when the number of the units that can be interfered with does not grow as quickly as the sample size. Ref. [134] develops the central limit theorem for direct and spillover effects under partial interference and stratified interference. Ref. [52] provides the central limit theorem for a peer effect under partial interference and stratified interference. However, under general interference, the asymptotic properties remain unsolved—even for the simplest difference in the means estimator. Second, interference becomes even harder to deal with when data complications are present. Refs. [120,121,135,136] consider noncompliance in an interference setting. Ref. [137] examines the censoring of time-to-event data in the presence of interference. However, for other data complications such as missing data and measurement error, no methods are yet available. Third, most of the literature focuses on the direct effect and the spillover effect. However, interference may be present in other settings, such as mediation analysis (see Ref. [138] for a mediation analysis under interference) and longitudinal studies, where different quantities are of interest. As a result, it is necessary to generalize the commonly used methods in these settings to account for the interference between units.

Compliance with ethics guidelines

Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang declare that they have no conflict of interest or financial conflicts to disclose.

[†] If the total number of units is N, then there are 2^N potential outcomes for each unit.

References

[1] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences. New York: Cambridge University Press; 2015.

[2] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61(4):962–73.

[3] Kuang K, Cui P, Li B, Jiang M, Yang S, Wang F. Treatment effect estimation with data-driven variable decomposition. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA; 2017.

[4] Athey S, Imbens GW, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. J R Stat Soc Ser B (Stat Methodol) 2018;80(4):597–623.

[5] Kuang K, Cui P, Li B, Jiang M, Yang S. Estimating treatment effect in the wild via differentiated confounder balancing. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017 Aug 13–17; Halifax, NS, Canada; 2017. p. 265–74.

[6] Imai K, Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. J Am Stat Assoc 2004;99(467):854–66.

[7] Egami N, Imai K. Causal interaction in factorial experiments: application to conjoint analysis. J Am Stat Assoc 2019;114(526):529–40.

[8] Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M. Causal effect inference with deep latent-variable models. In: Proceedings of Advances in Neural Information Processing Systems 30; 2017 Dec 4–9; Long Beach, CA, USA; 2017. p. 6446–56.

[9] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika 2009;96(1):187–99.

[10] Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. Am J Epidemiol 2019;188(1):250–7.

[11] Kuang K, Cui P, Athey S, Xiong R, Li B. Stable prediction across unknown environments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19–23; London, UK; 2018. p. 1617–26.

[12] Zhuang Y, Wu F, Chen C, Pan Y. Challenges and opportunities from big data to knowledge in AI 2.0. Front Inf Technol Elec Eng 2017;18(1):3–14.

[13] Pan Y. 2018 special issue on artificial intelligence 2.0: theories and applications. Front Inf Technol Elec Eng 2018;19(1):1–2.

[14] Hoerl C, McCormack T, Beck SR, editors. Understanding counterfactuals, understanding causation: issues in philosophy and psychology. New York: Oxford University Press; 2011.
[15] Pearl J, Glymour M, Jewell NP. Causal inference in statistics: a primer. Hoboken: John Wiley & Sons; 2016.
[16] Daniel RM, De Stavola BL, Vansteelandt S. Commentary: the formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? Int J Epidemiol 2016;45(6):1817–29.
[17] Pearl J. Causal and counterfactual inference. Forthcoming section in the handbook of rationality. Cambridge: MIT press; 2018.
[18] Goldfeld K. Considering sensitivity to unmeasured confounding: part 1 [Internet]. New York: Keith Golgfeld; 2019 Jan 2 [cited 2019 Jun 1]. Available from: https://www.rdatagen.net/post/what-does-it-mean-if-findings-are-sensitive-to-unmeasured-confounding/.
[19] Yule GU. Notes on the theory of association of attributes in statistics. Biometrika 1903;2(2):121–34.
[20] Simpson EH. The interpretation of interaction in contingency tables. J R Stat Soc B 1951;13(2):238–41.
[21] Chen H, Geng Z, Jia J. Criteria for surrogate end points. J R Stat Soc Series B Stat Methodol 2007;69(5):919–32.
[22] Geng Z, Liu Y, Liu C, Miao W. Evaluation of causal effects and local structure learning of causal networks. Annu Rev Stat Appl 2019;6(1):103–24.
[23] Pearl J. Is scientific knowledge useful for policy analysis? A peculiar theorem says: no. J Causal Infer 2014;2(1):109–12.
[24] Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125(7):605–13.
[25] Xu L, Pearl J. Structuring causal tree models with continuous variables. In: Proceedings of the Third Conference on Uncertainty in Artificial Intelligence. Arlington: AUAI Press; 1987. p. 170–9.
[26] Xu L. Deep bidirectional intelligence: alphazero, deep IA-search, deep IA-infer, and TPC causal learning. Appl Inf 2018;5(1):5.
[27] Xu L. Machine learning and causal analyses for modeling financial and economic data. Appl Inf 2018;5(1):11.
[28] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd ed. Cambridge: MIT Press; 2001.
[29] Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge University Press; 2000.
[30] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Appl Inform 2016;3(1):3.
[31] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7:2003–30.
[32] Zhang K, Hyvärinen A. On the identifiability of the post-nonlinear causal model. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; 2009 Jun 18–21; Montreal, QC, Canada. Arlington: AUAI Press; 2019. p. 647–55.
[33] Hoyer PO, Janzing D, Mooij JM, Peters J, Scholkopf B. Nonlinear causal discovery with additive noise models. In: Proceedings of International Conference on Neural Information Processing Systems; 2008 Dec 8–13; Vancouver, BC, Canada; 2008. p. 689–96.
[34] Zhang K, Hyvärinen A. Causality discovery with additive disturbances: an information-theoretical perspective. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J, editors. Machine learning and knowledge discovery in databases. Berlin: Springer; 2009. p. 570–85.
[35] Zhang K, Schölkopf B, Muandet K, Wang Z. Domain adaptation under target and conditional shift. In: Proceedings of the 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA; 2013. p. 819–27.
[36] Baroni P, Gabbay DM, Giacomin M, Van der Torre L. Handbook of formal argumentation. London: College Publications; 2018.
[37] Osborne J. Arguing to learn in science: the role of collaborative, critical discourse. Science 2010;328(5977):463–6.
[38] Shoham Y. Nonmonotonic reasoning and causation. Cogn Sci 1990;14(2):213–52.
[39] Liao B, Jin L, Koons RC. Dynamics of argumentation systems: a division-based method. Artif Intell 2011;175(11):1790–814.
[40] Sklar EI, Azhar MQ. Explanation through argumentation. In: Proceedings of the 6th International Conference on Human–Agent Interaction; 2018 Dec 15–18; Southampton, UK; 2018. p. 277–85.
[41] Fazzinga B, Flesca S, Furfaro F. Complexity of fundamental problems in probabilistic abstract argumentation: beyond independence. Artif Intell 2019;268:1–29.
[42] Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence; 2010 Jul 8–11; Catalina Island, CA, USA; 2000. p. 425–32.
[43] Kempthorne O. The design and analysis of experiments. New York: Wiley; 1952.
[44] Scheffe H. The analysis of variance. New York: John Wiley & Sons; 1959.
[45] Hinkelmann K, Kempthorne O. Design and analysis of experiments: volume 1: introduction to experimental design. 2nd ed. New York: John Wiley & Sons; 2007.
[46] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.
[47] Splawa-Neyman J. On the application of probability theory to agricultural experiments: essay on principles. Section 9. Stat Sci 1990;5(4):465–72.
[48] Ding P, Dasgupta T. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika 2018;105(1):45–56.
[49] Dasgupta T, Pillai NS, Rubin DB. Causal inference from 2^K factorial designs by using potential outcomes. J R Stat Soc Series B Stat Methodol 2015;77(4):727–53.
[50] Wu J, Ding P. Randomization tests for weak null hypotheses. 2018. arXiv:1809.07419.
[51] Miratrix LW, Sekhon JS, Yu B. Adjusting treatment effect estimates by post-stratification in randomized experiments. J R Stat Soc Series B Stat Methodol 2013;75(2):369–96.
[52] Li X, Ding P, Lin Q, Yang D, Liu JS. Randomization inference for peer effects. J Am Stat Assoc 2019:1–31.
[53] Li X, Ding P. General forms of finite population central limit theorems with applications to causal inference. J Am Stat Assoc 2017;112(520):1759–69.
[54] Zhao A, Ding P, Mukerjee R, Dasgupta T. Randomization-based causal inference from split-plot designs. Ann Stat 2018;46(5):1876–903.
[55] Mukerjee R, Dasgupta T, Rubin DB. Using standard tools from finite population sampling to improve causal inference for complex experiments. J Am Stat Assoc 2018;113(522):868–81.
[56] Fisher R. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.
[57] Freedman DA. On regression adjustments to experimental data. Adv Appl Math 2008;40(2):180–93.
[58] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. Ann Appl Stat 2013;7(1):295–318.
[59] Eicker F. Limit theorems for regressions with unequal and dependent errors. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967 Jun 21–Jul 18; Berkeley, CA, USA. Berkeley: University of California Press; 1967. p. 59–82.
[60] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967 Jun 21–Jul 18; Berkeley, CA, USA; Berkeley: University of California Press; 1967. p. 221–33.
[61] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 1980;48(4):817–38.
[62] Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Natl Acad Sci USA 2016;113(27):7383–90.
[63] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267–88.
[64] Lei L, Ding P. Regression adjustment in completely randomized experiments with a diverging number of covariates. 2018. arXiv:1806.07585.
[65] Ding P, Feller A, Miratrix L. Decomposing treatment effect variation. J Am Stat Assoc 2019;114(525):304–17.
[66] Lu J. Covariate adjustment in randomization-based causal inference for 2^K factorial designs. Stat Probab Lett 2016;119:11–20.
[67] Middleton JA. A unified theory of regression adjustment for design-based inference. 2018. arXiv:1803.06011.
[68] Cox DR. Randomization and concomitant variables in the design of experiments. In: Anderson TW, Styan GHP, Kallianpur GG, Krishnaiah PR, Ghosh JK, editors. Statistics and probability: essays in honor of CR Rao. Amsterdam: North-Holland; 1982. p. 197–202.
[69] Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. Ann Stat 2012;40(2):1263–82.
[70] Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment-control experiments. Proc Natl Acad Sci USA 2018;115(37):9157–62.
[71] Morgan KL, Rubin DB. Rerandomization to balance tiers of covariates. J Am Stat Assoc 2015;110(512):1412–21.
[72] Branson Z, Dasgupta T, Rubin DB. Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City department of education high school study. Ann Appl Stat 2016;10(4):1958–76.
[73] Li X, Ding P, Rubin DB. Rerandomization in 2^K factorial experiments. 2018. arXiv:1812.10911.
[74] Zhou Q, Ernst PA, Morgan KL, Rubin DB, Zhang A. Sequential rerandomization. Biometrika 2018;105(3):745–52.
[75] Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.
[76] Rubin DB. Comment on "randomization analysis of experimental data: the Fisher randomization test". J Am Stat Assoc 1980;75(371):591–3.
[77] Tukey JW. Tightening the clinical trial. Control Clin Trials 1993;14(4):266–85.
[78] Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. Stat Sci 2002;17(3):286–327.
[79] Aronow PM. A general method for detecting interference between units in randomized experiments. Sociol Methods Res 2012;41(1):3–16.
[80] Athey S, Eckles D, Imbens GW. Exact *p*-values for network interference. J Am Stat Assoc 2018;113(521):230–40.
[81] Basse G, Feller A, Toulis P. Exact tests for two-stage randomized designs in the presence of interference. 2017. arXiv:1709.08036.
[82] Ding P. A paradox from randomization-based causal inference. Stat Sci 2017;32(3):331–45.
[83] Rosenbaum PR. Exact confidence intervals for nonconstant effects by inverting the signed rank test. Am Stat 2003;57(2):132–8.
[84] Rigdon J, Hudgens MG. Randomization inference for treatment effects on a binary outcome. Stat Med 2015;34(6):924–35.
[85] Li X, Ding P. Exact confidence intervals for the average causal effect on a binary outcome. Stat Med 2016;35(6):957–60.
[86] Ding P, Li F. Causal inference: a missing data perspective. Stat Sci 2018;33(2):214–37.

[87] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat. Sci 1999;14:29–46.

[88] Greenland S, Pearl J. Adjustments and their consequences—collapsibility analysis using graphical models. Int Stat Rev 2011;79(3):401–26.

[89] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55.

[90] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952;47(260):663–85.

[91] Wright PG. Tariff on animal and vegetable oils. New York: Macmillan; 1928.

[92] Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour 1997;32 (3):441–62.

[93] Manski CF. Nonparametric bounds on treatment effects. Am Econ Rev 1990;80(2):319–23.

[94] Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Am Stat Assoc 1997;92(439):1171–6.

[95] Goldberger AS. Structural equation methods in the social sciences. Econometrica 1972;40(6):979–1001.

[96] Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Method 1994;23 (8):2379–412.

[97] Hernán MA, Robins JM. Causal inference. Boca Raton: Chapman & Hall; 2011.

[98] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91(434):444–55.

[99] Lin W, Feng R, Li H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. J Am Stat Assoc 2015;110(509):270–88.

[100] Kang H, Zhang A, Cai TT, Small DS. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. J Am Stat Assoc 2016;111(513):132–44.

[101] Wang L, Robins JM, Richardson TS. On falsification of the binary instrumental variable model. Biometrika 2017;104(1):229–36.

[102] Manski CF, Pepper JV. Monotone instrumental variables: with an application to the returns to schooling. Econometrica 2000;68(4):997–1010.

[103] Small DS. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. J Am Stat Assoc 2007;102(479):1049–58.

[104] Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika 2018;105(4):987–93.

[105] Miao W, Tchetgen Tchetgen E. Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. Am J Epidemiol 2017;185(10):950–3.

[106] Miao W, Tchetgen ET. A confounding cridge approach for couble negative control inference on causal effects. 2018. arXiv:1808.04945.

[107] Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21(3):383–8.

[108] Smith GD. Negative control exposures in epidemiologic studies. Epidemiology 2012;23(2):350–1.

[109] Flanders WD, Strickland MJ, Klein M. A new method for partial correction of residual confounding in time-series and other observational studies. Am J Epidemiol 2017;185(10):941–9.

[110] Rosenbaum PR. The role of known effects in observational studies. Biometrics 1989;45(2):557–69.

[111] Wang J, Zhao Q, Hastie T, Owen AB. Confounder adjustment in multiple hypothesis testing. Ann Stat 2017;45(5):1863–94.

[112] Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics 2012;13(3):539–52.

[113] Hong G, Raudenbush SW. Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. J Am Stat Assoc 2006;101(475):901–10.

[114] Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J Am Stat Assoc 2006;101 (476):1398–407.

[115] Halloran ME, Struchiner CJ. Causal inference in infectious diseases. Epidemiology 1995;6(2):142–51.

[116] Halloran ME, Struchiner CJ. Study designs for dependent happenings. Epidemiology 1991;2(5):331–8.

[117] Halloran ME, Hudgens MG. Dependent happenings: a recent methodological review. Curr Epidemiol Rep 2016;3(4):297–305.

[118] Hudgens MG, Halloran ME. Toward causal inference with interference. J Am Stat Assoc 2008;103(482):832–42.

[119] Basse G, Feller A. Analyzing two-stage experiments in the presence of interference. J Am Stat Assoc 2018;113(521):41–55.

[120] Forastiere L, Mealli F, VanderWeele TJ. Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using bayesian principal stratification. J Am Stat Assoc 2016;111 (514):510–25.

[121] Kang H, Imbens G. Peer encouragement designs in causal inference with partial interference and identification of local average network effects. 2016. arXiv:1609.04464.

[122] Rigdon J, Hudgens MG. Exact confidence intervals in the presence of interference. Stat Probab Lett 2015;105:130–5.

[123] Aronow PM, Samii C. Estimating average causal effects under interference between units. 2018. arXiv:1305.6156v4.

[124] Aronow PM, Samii C. Estimating average causal effects under general interference, with application to a social network experiment. Ann Appl Stat 2017;11(4):1912–47.

[125] Choi D. Estimation of monotone treatment effects in network experiments. J Am Stat Assoc 2017;112(519):1147–55.

[126] Forastiere L, Airoldi EM, Mealli F. Identification and estimation of treatment and interference effects in observational studies on networks. 2016. arXiv:1609.06245.

[127] Eckles D, Karrer B, Ugander J. Design and analysis of experiments in networks: reducing bias from interference. J Causal Inference 2017;5(1): 1–23.

[128] Eckles D, Kizilcec RF, Bakshy E. Estimating peer effects in networks with peer encouragement designs. Proc Natl Acad Sci USA 2016;113(27):7316–22.

[129] Jagadeesan R, Pillai N, Volfovsky A. Designs for estimating the treatment effect in networks with interference. 2017. arXiv:1705.08524.

[130] Bowers J, Fredrickson MM, Panagopoulos C. Reasoning about interference between units: a general framework. Polit Anal 2013;21(1):97–124.

[131] Toulis P, Kao E. Estimation of causal peer influence effects. In: Proceedings of 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA; 2013. p. 1489–97.

[132] Basse GW, Feller A, Toulis P. Randomization tests of causal effects under interference. Biometrika 2019;106(2):487–94.

[133] Sävje F, Aronow PM, Hudgens MG. Average treatment effects in the presence of unknown interference. 2017. arXiv:1711.06399.

[134] Liu L, Hudgens MG. Large sample randomization inference of causal effects in the presence of interference. J Am Stat Assoc 2014;109(505):288–301.

[135] Imai K, Jiang Z, Malani A. Causal inference with interference and noncompliance in two-stage randomized experiments. Technical report. Princeton: Princeton University; 2018.

[136] Kang H, Keele L. Spillover effects in cluster randomized trials with noncompliance. 2018. arXiv:1808.06418.

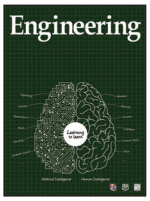
[137] Loh WW, Hudgens MG, Clemens JD, Ali M, Emch ME. Randomization inference with general interference and censoring. 2018. arXiv:1803.02302.

[138] Vanderweele TJ, Hong G, Jones SM, Brown JL. Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. J Am Stat Assoc 2013;108(502):469–82.

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Artificial Intelligence—Review

因果推理 †

况琨 [a,*]，李廉 [b]，耿直 [c]，徐雷 [d]，张坤 [e]，廖备水 [f]，黄华新 [f]，丁鹏 [g]，苗旺 [h]，蒋智超 [i]

[a] *College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China*
[b] *Department of Computer Science and Technology, HeFei University of Technology, Hefei 230009, China*
[c] *School of Mathematical Science, Peking University, Beijing 100871, China*
[d] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[e] *Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[f] *School of Humanities, Zhejiang University, Hangzhou 310058, China*
[g] *University of California Berkeley, Berkeley, CA 94720, USA*
[h] *Guanghua School of Management, Peking University, Beijing 100871, China*
[i] *Department of Government and Department of Statistics, Harvard University, Cambridge, MA 02138, USA*

ARTICLE INFO

Article history:
Received 8 May 2019
Revised 31 July 2019
Accepted 26 August 2019
Available online 8 January 2020

关键词
因果推断
指示变量
阴性对照
因果推理和解释
因果发现
反事实推理
治疗效果评价

摘要

因果推理是解释性分析的强大建模工具，它可使当前的机器学习变得可解释。如何将因果推理与机器学习相结合，开发可解释人工智能（XAI）算法，是迈向人工智能2.0的关键步骤之一。为了将因果推理的知识带给机器学习和人工智能领域的学者，我们邀请从事因果推理的研究人员，从因果推理的不同方面撰写了本综述。本综述包括以下几个部分：况琨博士的"平均因果效应评估——简要回顾与展望"，李廉教授的"反事实推理的归因问题"，耿直教授的"Yule-Simpson悖论和替代指标悖论"，徐雷教授的"因果发现CPT方法"，张坤教授的"从观测数据中发现因果关系"，廖备水和黄华新教授的"形式论辩在因果推理和解释中的作用"，丁鹏教授的"复杂实验中的因果推断"，苗旺教授的"观察性研究中的工具变量和阴性对照方法"，蒋智超博士的"有干扰下的因果推断"。

1. 平均因果效应评估——简要回顾及展望

机器学习方法已在许多领域取得了巨大成功，但其中大部分都缺乏可解释性。因果推理是一种强有力的建模工具，可用于解释性分析，其可能使当前的机器学习能够做出可解释的预测。本文回顾了两个用于估计因果效应的经典算法，并讨论了实际应用中因果效应评估存在的挑战。此外，我们提出了一种可能的方法，通过将因果推理与机器学习相结合来开发可解释的人工智能（explainable artificial intelligence, XAI）算法。

1.1. 问题和符号

基于潜在的结果模型[1]，我们研究的问题是如何准确地评估治疗变量的因果效应。对于每个样本i（$i = 1, 2, \ldots, n$），我们观测到其治疗变量T_i、结果变量Y_i和特征变量X_i。基于治疗变量的取值（$T = 1$和$T = 0$），结果变量

† The authors contributed equally to this work. The symbol definitions and notations of each section are relatively independent.

* Corresponding author.
E-mail address: kunkuang@zju.edu.cn (K. Kuang)

2095-8099/© 2020 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文：Engineering 2020, 6(3): 253–263
引用本文：Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, Zhichao Jiang. Causal Inference. *Engineering*, https://doi.org/10.1016/j.eng.2019.08.016

存在两个潜在结果$Y_i(1)$和$Y_i(0)$。实际观测到的结果Y_i^{obs}可表示为：

$$Y_i^{\mathrm{obs}} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0) \quad (1)$$

基于样本的潜在结果，我们可以定义治疗变量的评估因果效应为：

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] \quad (2)$$

也可以定义在治疗样本（$T = 1$）上的平均因果效应为：

$$\tau_{\mathrm{t}} = \mathbb{E}[Y_i(1) - Y_i(0) | T_i = 1]$$

为了准确评估因果效应τ和τ_{t}，我们假设无混淆性$T_i \perp [Y_i(1), Y_i(0)] | \boldsymbol{X}_i$和重叠性$0 < p(T_i = 1 | \boldsymbol{X}_i) < 1$。

1.2. 两类因果评估算子

在这里，我们简要介绍评估因果效应的两类常用方法，并讨论它们面对高维变量的扩展和应用。

1.2.1. 倾向值倒数加权

在完全随机的实验中，治疗随机分配到样本，意味着$T_i \perp \boldsymbol{X}_i$。然而，在观测数据中，治疗$T_i$是基于样本特征$\boldsymbol{X}_i$指定的。为了消除$\boldsymbol{X}_i$导致的混杂效应，倾向值表示为$e(\boldsymbol{X}_i) = (T_i = 1 | \boldsymbol{X}_i)$来加权样本。基于倾向值，$\tau$可以通过以下方式估算：

$$\tau = \mathbb{E}\left[\frac{Y_i^{\mathrm{obs}} T_i}{e(\boldsymbol{X}_i)} - \frac{Y_i^{\mathrm{obs}}(1 - T_i)}{1 - e(\boldsymbol{X}_i)}\right] \quad (3)$$

通过结合倾向加权和回归，我们还可以用双稳健方法估计治疗效果[2]。在高维情况下，并非所有观察到的变量都是混淆变量。为了解决这个问题，Kuang等[3]建议将所有观察到的变量分为两部分：混淆变量用于评估倾向值，调整变量用于减少估计因果效应方差。

1.2.2. 混淆变量平衡

消除混淆偏差的另一种常用的方法是通过直接样本加权来平衡治疗组（$T = 1$）和对照组（$T = 0$）之间混淆变量的分布，并将τ_{t}估计为：

$$\tau_{\mathrm{t}} = \mathbb{E}\left[Y_i^{\mathrm{obs}} | T_i = 1\right] - \mathbb{E}\left[W_j Y_i^{\mathrm{obs}} | T_j = 0\right] \quad (4)$$

式中，样本权重W可以通过混淆变量直接平衡[4]学习得到，如下：

$$W = \arg\min_{W} \left\| \mathbb{E}\left[Y_i^{\mathrm{obs}} | T_i = 1\right] - \mathbb{E}\left[W_j Y_i^{\mathrm{obs}} | T_j = 0\right] \right\|_{\infty}^{2} \quad (5)$$

在高维情况下，不同的混淆变量可能导致不同的混淆偏差。因此，Kuang等[5]建议联合学习混淆变量权重用于区分不同的混淆变量，以及样本权重用于平衡混淆变量分布，并提出混淆变量区分性（differentiated confounder balancing, DCB）算法来评估因果效应。

1.3. 存在的挑战

最近，很多有效的方法用于在观察性研究中评估因果效应，但如何使这些方法在实践中变得有用仍然存在许多挑战，主要挑战表现在以下几个方面。

1.3.1. 治疗变量从二值到连续

现有算法主要用于评估二值治疗变量的因果效应，并在实际中实现良好的性能。但是在许多实际应用中，我们不仅关心二值治疗变量（治疗与否）的因果效应，更关心连续治疗变量的因果效应。

1.3.2. 多维治疗变量的相互作用

实际上，治疗可以由多个变量及其相互作用组成。在社交营销中，人们可能对不同广告策略的综合因果效应感兴趣。在支持治疗组合的因果分析方面还需要做更多的工作。

1.3.3. 未观察到的混淆变量

未观察到的混淆变量等同于违反非混淆性假设，并且它是不可测试的。控制高维变量可能使非混淆性假设更加合理，但对倾向值估计和混淆变量平衡提出了新的挑战。

1.3.4. 重叠性假设的限制

尽管重叠性假设是可测试的，但它在实践中会引起很多问题，包括如何检测变量分布中是否缺乏重叠，以及如何处理这种缺陷，特别是在高维度环境中。此外，估计因果效应仅适用于重叠性假设成立的样本区域。

为了解决上述挑战，最近，很多相关工作和算法被相继提出，包括连续治疗变量[6]、治疗变量的相互作用[7]、未观测到的混淆变量[8]以及重叠性假设的限制[9,10]。

1.4. 走向因果和稳定的预测

大多数预测算法缺乏可解释性，这使得它们在许多实际应用中缺乏吸引力，特别是那些需要决策的应用。此外，

大多数现有机器学习算法都是关联驱动的，这导致它们在测试数据中的性能不稳定，因为测试数据的分布可能与训练数据不同。因此，开发可解释的并对来自未知测试数据的分布变化保持稳定预测的预测算法非常有用。

假设因果关系在数据集之间是不变的，那么合理的方法是探索因果和稳定预测的因果知识来实现可解释的稳定预测。受到来自因果推理文献中混淆变量平衡技术的启发，Kuang等[11]提出了实现因果和稳定预测的可能解决方案。他们提出了一个全局变量平衡正则约束项，以隔离每个预测变量的影响，从而恢复每个预测变量和结果变量之间的因果关系，用于指导机器学习算法的学习，实现在未知数据集中进行可解释的稳定预测。

总体而言，如何将因果推理与机器学习深入结合以开发可解释的人工智能算法是通往第二代人工智能[12,13]的关键，目前还存在很多问题、挑战和机遇。

2. 反事实推理的归因问题

本节中，原因变量x和结果变量y都是二值的。

反事实推理是因果推理的重要部分，简单地说，反事实推理是在事件x已经出现（$x = 1$），并且事件y发生（$y = 1$）的前提下，反过来推理如果事件x不出现（$x = 0$），则事件y不发生（$y = 0$）的概率，用公式表示为：

$$P(y_{x=0} = 0|x = 1,\ y = 1) \qquad （6）$$

式中，$y_{x=0}$是反事实推理的一个记号，表示如果$x = 0$时y的取值，与条件概率$P(y|x = 0)$是不同的概念。这个公式反映了没有事件x则没有事件y的概率，即原因的必要性（x作为y的原因）。这在社会科学或者逻辑科学中称为归因问题（attribution problem），在法律学上称为“若无准则”（but-for criterion）。归因问题已经有了比较长的研究历史了，但是以往的方法主要是社会科学的方法，如案例调查、统计分析、实验设计等，基本是定性的，且依赖于经验和直觉。随着大数据的出现，数据驱动的研究归因问题的定量化方法出现了，从而使得推理过程更加科学与合理。

归因问题另一个孪生的说法：在原因x未出现（$x = 0$），且事件y也未发生（$y = 0$）的前提下，如果原因x出现，那么事件y出现的概率为：

$$P(y_{x=1} = 1|x = 0,\ y = 0) \qquad （7）$$

该公式反映了原因x导致事件y发生的概率，即原因的充分性。

在基于数据的推理研究中，原因的必要性和充分性是因果关系的不同侧面，尽管在计算公式上有所不同，但其基本精神是一致的。

反事实推理与人类的反思行为相对应，反思是智能活动的重要特征。推理使得人们在采取某个动作时，预测相应的结果，而反思却是在已有的结果面前，思考如何改进结果。反思虽然不能改变已经发生的结局，却可以为以后的行为提供修正，是用过去的知识指导未来行动的数学模型，只有具备了反思能力的智能才称得上是真正的智能。

反思在人类的日常生活中也是很重要的，例如，张某和李某同时做了癌症手术，张某又接受了放化治疗，结果两人都得到康复。因此张某就会反思，如果不做放化治疗，是否也能得到康复，显然我们不能因为李某的康复，就认为张某不做放化治疗也能康复。这类问题在医疗、法院审判等场合是大量存在的。我们关心对于具体的个案而言，当结果已经发生时，其原因究竟是什么。这时，一般的统计数据，如放疗成功率、交通事故率等并不能说明问题。通过归因推理计算某些原因的必要性在这些领域有着关键的意义[14]。

社会科学研究中有一个很有影响的关于因果分析的理论，即澳大利亚哲学家Mackie于20世纪60年代提出的INUS理论，INUS是Insufficient but Necessary part of a Unnecessary but Sufficient condition的缩写，意思是某个充分不必要条件下的必要不充分部分[2]。INUS理论认为，在事实集合$\boldsymbol{X}$已经导致事实y发生的前提下（即$\boldsymbol{X}$对于y是充分却可能不必要的），我们认为$\{x_1, x_2, \ldots, x_k\} \subseteq \boldsymbol{X}$是$y$的原因，如果$\{x_1, x_2, \ldots, x_k\}$是$\boldsymbol{X}$的必要却可能不充分部分。过去这一理论的语义上有很多不确定的成分，所用的方法是定性的，或者经验主义的，没有可行的算法用于量化计算。而在基于大数据的推理技术出现以后，这一理论赋予了准确的含义，并且可以通过算法进行定量描述。

关于式（6）的计算形成了反事实推理中的一个重要研究内容，目前还没有通用的计算该公式的方法。在实际问题中，引进一个很多情况下都能满足的所谓“单调性”假设：

$$y_{x=1} \geqslant y_{x=0}$$

单调性的直观意思是：采取某种措施（$x = 1$）后的效果y总不会低于未采取措施（$x = 0$）的效果。例如，在流行病学中，一个人不会在未采取隔离措施（$x = 0$）时健康（$y = 1$），在采取隔离措施后（$x = 1$）反被感染（$y = 1$）。由于单调性，式（6）通过推导得到：

$$\begin{aligned} P(y_{x=0} = 0|x = 1,\ y = 1) &= \frac{P(y = 1) - P(y_{x=0} = 1)}{P(x = 1,\ y = 1)} \\ &= \frac{P(y = 1|x - 1) - P(y = 1|x = 0)}{P(y = 1|x - 1)} \\ &\quad + \frac{P(y = 1|x = 0) - P(y_{x=0} = 1)}{P(y = 1|x = 1)} \end{aligned} \qquad (8)$$

式（8）分为两项，前面一项是在风险统计中熟悉的归因风险部分（attributable risk fraction），或者也叫额外风险率（excess risk ratio），它反映了在$x = 1$和$x = 0$不同的措施下的风险率。而后一项就是特别要提出的混杂影响部分（confounding affect fraction），它反映了其他变量干扰的影响。$P(y_{x=0} = 1)$称为do-操作，有时也写作$P(y = 1|\mathrm{do}(x = 0))$。这是在实验（操纵）条件下，将其他变量固定，而单独考察$x = 0$时$y = 1$的概率。而条件概率$P(y = 1|x = 0)$表示自然条件下，$x = 0$时$y = 1$的概率（其他变量的值不作限制），$P(y_{x=0} = 1)$和$P(y = 1|x = 0)$有着不一样的含义。在do-操作（实验条件下）时，我们看到的是x与y之间单纯的（因果）关系。而在自然条件下，y的变化来自两个方面，一个是x的变化直接引起的，另一个是通过其他变量间接引起的，这个现象称为混杂。两者之间的差$P(y = 1|x = 0) - P(y_{x=0} = 1)$表示了混杂的程度。混杂现象干扰了对于真正原因$x$的计算。在有些情况下，$x$的变化的确引起了$y$的变化，但$x$可能根本不是$y$变化的原因（如鸡叫之后太阳升起）。虽然我们可以通过科学实验来排除混杂，以便找出引起y变化的真正原因，但是在许多社会科学问题研究中，包括一些自然科学问题研究中，科学的实验是很难实施的，甚至是不可能的。我们手里能够得到的数据只有观察数据，因此如何从观察数据中识别混杂，以找出真正的因果关系是人工智能一项重要的研究内容。

为了更加具体地说明归因部分和混杂部分之间的关系，以及它们对于归因问题（原因的必要性）所起的作用，我们引用文献[15]中的例子。在这个例子中，赵某由于疼痛，去药店买了某种止痛药服用，结果死了。官司打到法院，要求药厂承担部分责任。药厂和原告律师分别出具了药品的检验结果（experimental）和市场调查结果（nonexperimental），如表1所示，其中，$x = 1$表示吃药，$y = 1$表示死亡。

药厂的数据来源于严格的药品安全实验标准，而律师的数据来源于市场调查，在患病的人里面根据自愿服药进行统计。药厂的理由是：药品已经通过检验，虽然吃药的死亡概率有所提高（由0.014提高到0.016），但是比起止痛效果而言，这点提高还是可以接受，并且符合药品上市的规定，因此根据传统的风险归因计算（额外风险率），药厂承担的责任是：

$$\frac{P(y = 1|x = 1) - P(y = 1|x = 0)}{P(y = 1|x = 1)} = \frac{0.016 - 0.014}{0.016} = 0.125 \qquad (9)$$

律师的理由是：药品的检验是在随机选择的条件下进行的，并没有征求受试者本人的意愿，因此实验是有偏置的（bias），并不符合实际服药情况，而市场调查的观察数据则是完全自愿的，未受到任何干预。从数据中看出，观察数据与实验数据相差很大。在赵某已经死亡的前提下，药厂的责任（即如果不吃药就会不死亡的概率）应该按照反事实的公式计算，其结果为：

$$\begin{aligned} &\frac{P(y = 1|x - 1) - P(y = 1|x = 0)}{P(y = 1|x - 1)} \\ &+ \frac{P(y = 1|x = 0) - P(y_{x=0} = 1)}{P(y = 1|x = 1)} \\ &= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = 1 \end{aligned} \qquad (10)$$

因此药厂应对赵某的死亡负全责。

表面上来看，对于市场调查的数据，吃药而死的只占0.2%，而不吃药死亡的占了2.8%，非常有利于药厂（药厂的风险责任是–13，比实验数据还要好），但是仔细分析后，由于混杂成分占了$P(y = 1|x = 0) - P(y_{x=0} = 1) = 0.014$，也就是有一半的人，不是因为未吃药而死亡，是因为其他原因死亡，因此这部分的功劳不能归因于药品，由此提升了药厂的责任。当然，在这个例子中，药厂的责任是否真

表1　药品诉讼案例的检验结果和市场调查结果

Outcomes	Experimental data (number of patients)		Non-experimental data (number of patients)	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$
Deaths (y =1)	16	14	2	28
Survivals (y =0)	984	986	998	972

的就是100%，以及计算公式的合理性与科学性还有一些质疑[16]。但是，这个案例说明在观察数据中存在的混杂变量会干扰真正的因果发现，而如何有效地识别混杂现象是因果推理中的现实问题，也是反事实推理中具有实际意义的问题。

在数据科学里，数据包括人为生成的数据和客观产生的数据。在客观产生的数据中又有实验数据和观察数据，前者是在实验条件下所搜集的数据，后者是在自然条件下搜集的数据。观察数据虽然客观、易于获取、成本较低，但是其中的混杂问题往往构成了因果推理的障碍[17]。特别在客观世界中可能还会有未知的变量（即隐变量），这些变量我们未能观察到，同样可以对已知的变量发生作用，这种情况称为非测定混杂。或者说，已知变量对于隐变量带来的非测定混杂可能是敏感的，目前的研究还处于十分初步的阶段，希望了解更多细节的读者可参考文献[18]。

3. Yule-Simpson 悖论和替代指标悖论

两个变量的相关性可能会由于忽略了第三个变量而发生非常大的变化，甚至可能从正相关变为负相关，这个现象称为Yule-Simpson悖论[19,20]。第三个变量是一个混杂因素。表2给出了一个数值例子。风险差（risk difference, RD）定义为吸烟组患肺癌比率与非吸烟组患肺癌比率的差值：RD = (80/200) – (100/200) = –0.10，其值为负。表3给出了将这400人按照性别分组的结果。可以看到按性别分组后，发生了很大变化，男性组和女性组的风险差都变成正的，即0.10。这意味着吸烟分别对男性和对女性都有害，但是吸烟对人类有益。

与相关性推断的关键区别在于，因果推断必须考虑是否可能存在影响处理变量和结果变量的公共原因，称为混杂因素。混杂因素可能导致推断因果作用时产生混杂偏倚。在试验性研究中，我们可以设计试验方案，根据某些变量的水平制定处理或暴露的分配概率；在随机化试验中，给每个个体随机地分配处理和暴露，没有影响处理分配的变量，因此不存在混杂因素。在观察性研究中，为了进行因果推断，需要观测充分多的混杂因素，或者观测一个独立所有混杂因素的工具变量。可是，利用观测到的数据不能确认是否观测到了充分多的混杂因素或工具变量。根据观察性研究得到的数据进行因果推断时要求一些假定，这些假定是不能用数据检验的。

表2 吸烟与肺癌的关系

Condition	Number of persons		
	Cancer	No cancer	Total
Smoking	80	120	200
No smoking	100	100	200

表3 吸烟与肺癌的关系：按照性别的分类

Condition	Males		Females	
	Cancer	No cancer	Cancer	No cancer
Smoking	35	15	45	105
No smoking	90	60	10	40

当感兴趣的终点指标不易观测时，取而代之观测一个替代指标（如生物标记物），然后用替代指标的因果作用预测处理未观测终点指标的因果作用。目前，已经有各种选择和确定替代指标的准则。但是，这些准则难以避免替代指标悖论，即处理对替代指标有正的因果作用，而且替代指标对终点指标也有正的因果作用，但是，处理对终点指标反而有负的因果作用[21]。文献[21,22]给出了替代指标悖论的数值例子。这个悖论也质疑了科学知识是否对决策分析有用[23]。有一个著名的实例，医生知道心律失常是猝死的危险因素，因此，临床试验将心律失常作为猝死的替代指标。但是，一些能有效纠正心律失常的药物，后来发现不但不能减少猝死，反而导致数万人过早死亡[24]。如何选择和确定替代指标的准则还有待于进一步研究。

Yule-Simpson悖论和替代指标悖论告诉我们：从数据得到的结论可能会被未观测的混杂因素逆转。这两个悖论还强调了获取数据方法的重要性。首先，随机化试验是因果推断的金标准。其次，如果不能采用随机化试验，试验设计需要试图平衡处理组和对照组的混杂因素。再次，可以考虑鼓励试验方法，随机地鼓励一部分人接受处理或避免暴露，使得这些鼓励能够改变他们处理和暴露的概率，实际上这种鼓励试验方法设计了一个工具变量“鼓励”。最后，对于纯观察性研究，我们不得不根据专业知识论证因果推断所要求的假定，并且采用敏感性分析探讨偏离这些假定的情况下影响因果推断结论的程度。这两个悖论还指出了三段式推理和传递性推理也许不能应用于统计得到的总体结论。统计得到的总体结论也许会出现这样的现象：吸烟对男性和女性都有害，尽管任何一个人不是男人就是女人，但是，吸烟对人也可能有益；药能纠正心律失常，纠正心律失常可以延长寿命，但是，该药也许会缩短寿命。

4. 因果发现 CPT 方法

如何从观察数据中发现和分析因果结构，近百年来研究者已做了广泛努力。大致都沿着一个类似的思维方向。首先，对欲观察的多个变量先假设一个因果结构，它直接或间接地描述这些变量的概率分布，而变量间关系反映因果关系。最简单的情况是一个单变量随机模型$x \rightarrow y$，而复杂情况下为一个有向无环图（directed acyclic graph, DAG），并满足某些约束条件。然后，通过观察数据学习模型中的未知参数，监测模型能否最佳描述观察数据，并满足主要的约束条件，从而检验模型和因果方向的正确性。常用方法包括Rubin因果模型、结构方程模型、函数因果模型、非线性可加噪声模型、线性非正态无环模型、后非线性模型、型结构发现模型[25]和ρ图方程模型[26]。

最近提出的因果势理论（causal potential theory, CPT）源于很不同的思维[27]。回归物理学，因果关系被视为由某种势能引起的内在动力学性质。环境U中一对变量x和y之间的因果关系，并不简单的是这对变量间的局部关系，而是全局趋势在二元关系上的投影。无需事先假定因果结构，从x和y的样本估计非参数分布$p(x, y|U)$，再通过吉布斯分布获得对应的因果势能$E(x, y|U) \propto -\ln p(x, y|U)$。发生在$x$和$y$处的事件，源于动力学$[\dot{x}_t, \dot{y}_t] \propto [g_x, g_y]$，这里记号$g_u \triangleq \nabla_u E, \dot{u}_t \triangleq \mathrm{d}u/\mathrm{d}t$。该动力流推向能量最低处或可能发生事件的区域。如表4所示，可以通过对g_y和x之间以及g_x和y之间的独立性检验判定因果方向，或者无向，也可以近似检验二阶独立，即相关系数是否为0。两个检验可用多元检验方法联合进行。

估算$p(x, y|U)$也可不用非参数方法，而是通过某种参数模型，还可通过x和y的样本直接估算$[g_x, g_y]$，尤其是考虑在环境U下如何估算$p(x, y|U)$。进一步发展是探讨估计多个变量分布和多个变量之间的因果结构。下面介绍沿两个方向的可能发展。

（1）回顾PC算法[28]。通过独立和条件独立检验，发现DAG因果结构，其中考虑如何集成表4的CPT检验。当PC算法要剪的边与表4中结论冲突时，若x, y在第三变量条件下独立，则不剪边；即使x, y独立也不急于剪枝，可判为弱定向。

（2）分三步进行TPC学习[26]。一是用PC算法或上述集成，获得DAG的Topology。二是考虑变量间相关系数ρ进行路径分析，得到所谓的ρ图方程组，如文献[26]中的式（29）和式（33），形成一个变量约束在[–1,+1]间的多项式方程组，然后用Wu的方法来求解所有未知的ρ参数。若方程组有唯一或有限个解，表示对应的DAG有全局且定量的支撑。若没有解，则加入新的边或变量；若有无限多解，则剪掉某边或变量。三是用CPT（表4）确定每个边的Casual方向。然后，对于剩下的未能定向的边，再用类似PC的算法对v节点进行处理。

5. 从观测数据中发现因果关系

因果性是科学中的一个基本的概念，它在提供解释、预测以及决策和控制中扮演着重要的角色[28,29]。现代因果关系的研究中，有两类本质的问题需要解决。一类大致称为“因果推断”——假设已有部分或完全的因果结构，如何估计一个变量对另一个变量的因果影响的大小？关于这类研究，有兴趣者可参考文献[29]及其所引文章。该类问题常假设因果结构已知，但我们怎样才能知道因果关系？为找到因果关系，传统的方式是求助于人为干涉或者随机试验，但这在很多情况下太昂贵、太费时，甚至现实中不可行。因此，越来越多的人开始重视“因果发现”——通过分析被动观测的数据找出背后的因果关系。在过去30年，因果发现这一领域取得了很大进展，这部分归功于计算机技术的进步。这些进步包括收集、存储大数据的能力以及计算速度的提高。在一些领域，我们需要用到天气卫星图像、核磁功能扫描图像，或基因表达数据，变量的个数可达百万之多，而且一般情况下用来减小因果假设搜索空间的背景知识是很少的。若没有自动搜索，因果发现的实用性会极其受限。越来越快的计算机，以及大的内存和存储空间，使得因果发现的自动搜索算法可以处理大规模的实际问题。

统计学中有一个脍炙人口的说法——因果性蕴含了相关性，但相关性并没有蕴含因果性。我们觉得后半句改

表4 CPT因果关系分析的两条路

$\nabla_U E_U$	$y \rightarrow x$		$x \rightarrow y$		$x \perp\!\!\!\perp y$		$x?y$	
	Road_A	Road_B	Road_A	Road_B	Road_A	Road_B	Road_A	Road_B
g_x	dependent of y	$\xi(x, y) + \varepsilon$	$\perp\!\!\!\perp y$	$\xi(x) + \varepsilon$	$\perp\!\!\!\perp y$	$\xi(x) + \varepsilon$	dependent of y	$\xi(x, y) + \varepsilon$
g_y	$\perp\!\!\!\perp x$	$\eta(y) + \varepsilon$	dependent of x	$\eta(x, y) + \varepsilon$	$\perp\!\!\!\perp x$	$\eta(y) + \varepsilon$	dependent of x	$\eta(x, y) + \varepsilon$

成“但相关性并不直接蕴含因果性”会更公允。事实上，现在已经清楚地看到，在适当的假设下，通过分析观测数据可以找出部分或者完整的因果结构信息（通常用有向图表示）。从20世纪90年代开始，数据中的条件独立性已被用作约束条件来重构因果信息。这类基于约束的方法包括PC (Peter-Clark)算法和快速因果推理（fast causal inference, FCI）算法[28]。如果假设给定系统没有混淆因子（混淆因子的定义是给定系统中两个变量的未观测到的直接的因），那么PC算法的结果是渐进正确的。即使存在混淆因子，FCI算法的结果亦渐进正确。只要有可靠的条件独立检验的方法，这类算法可以处理各种因果关系和数据分布，所以有较广的实用性。但同时，它们的结果往往并不包含全部的因果信息——它们的输出结果是（独立）等价类。作为一个集合，这个类包含了所有具有同样的（条件）独立关系的因果结构。若假设没有混淆因子，还存在结果渐进正确的基于分数的方法，这些方法通过优化恰当的分数来搜索因果结构。这类方法中，贪婪等价类搜索（greedy equivalence search, GES）可直接在等价类空间进行搜索，因此已经得到很多应用。

在过去13年间，人们更进一步地发现基于恰当定义的函数因果模型可用来区分等价类中不同的因果结构。这个进展归功于关于因果机制的额外假设。函数因果模型把果变量Y写成直接的因X和噪声E的一个函数，数学描述是$Y=f(X,E)$，这里E和X是相互统计独立的。如果在f上没有任何约束条件，那么对于任意两个给定的变量，其中一个总能被写成另一个变量以及与之独立的噪声的函数。如此一来，因果的不对称性就没法得以体现[30]。幸运的是，如果我们恰当地约束函数类，就能找到X和Y之间的因果方向。这是因为若在错误的方向上估计因果模型，估计出的噪声和假设的因之间不可能统计独立，而在正确的方向上它们是独立的。这些函数因果模型类包括：

（1）线性非正态无环模型（linear non-Gaussian model, LiNGAM）[31]：它假设因果关系是线性的，而噪声是非正态的。

（2）后非线性模型（post-nonlinear, PNL）[32]：它考虑了因的非线性影响以及经常存在的测量过程非线性变形。

（3）非线性可加噪声模型（additive noise model, ANM）[33,34]：它描述了因的非线性影响并假设噪声是可加的。

若对如何用这些方法区分因果以及如何用它们从多个变量找到因果图感兴趣，可参考文献[30]。

因果发现是通过分析观测到的数据实现的。这些数据是由背后的因果过程以及观测和采样过程产生的。因此，在解决实际问题时，我们需要考虑因果过程以及观测过程带来的挑战。比如，从神经心理学中常用的血糖依赖水平（blood-oxygenation-level-dependent, BOLD）时间序列来发现背后的因果过程并不容易，部分由以下原因导致：因果交互可能是非线性的；数据的采样率跟背后的动态过程相比太低了；因果模型中可能存在反馈回路；过程具有非平稳性；可能存在混淆因子。在临床研究中，我们常常有很多缺失数据。网上收集的数据或者医院采集的数据一般都有选择偏差。有些数据集里面同时有类别变量（或离散变量）以及连续变量，这可能让条件独立检验以及寻找合适的函数因果模型类别变得困难。近年来这些问题基本都已引起注意，也出现了一些相应的处理方法。

机器学习的发展已经促进了因果发现的研究，因为机器学习为从数据中找寻信息提供了基本工具。另一方面，因果信息描述了过程的性质，从而提供了关于数据分布的一系列约束条件，而这些约束条件可帮助我们更好地理解和解决数据分布有变化时的机器学习问题。特别是若要从异构数据中学习有用信息，我们很自然地需要学习数据异构的性质，并为之建模，而这一步可受益于因果模型。这类问题包括领域适应（或迁移学习）[35]、半监督学习，以及从正类和无标记样本中学习。最近几年，利用因果模型帮助建立推荐系统以及进行强化学习也慢慢引起重视。

6. 形式论辩在因果推理和解释中的作用

在本节中，我们将以概述形式讨论论辩为何以及如何在因果推理和解释中发挥重要作用。形式论辩通过构造论证、比较论证和评估论证来实现推理[36]。论证通常由一个主张以及支持该主张的前提组成。前提可以是观察信息、假设或其他论证的中间结论。主张、前提和它们之间的推理关系都有可能受到攻击[37]。当一个论证能抵御所有攻击时，它才能够被接受。在人工智能领域中，形式论辩是建模可废止推理的一般形式，它为证明和解释因果关系提供了一种自然的方式。形式论辩也是机器学习的补充，可用于学习、推理和解释因果关系。

6.1. 非单调性和可废止性

因果推理是确定因果关系的过程。因果关系（即原因和结果之间的关系）通常是可废止和非单调的。一方面，因果推理规则通常是可废止的。因果规则可以用“c引起

e”来表示，其中，e是某种结果，c则是一个可能的原因。因果联结词并不是实质蕴涵的，而是带有强度和可能性的可废止条件式。例如，“转动点火开关是引起发动机启动的原因，但这并不意味着发动机一定会启动，因为发动机启动还和其他因素相关，如是否有电池，电池是否有电，是否有燃气，等等”[38]。另一方面，因果推理是非单调的。这意味着我们只能暂时得到因果关系，当我们得到更多的信息时，之前所获得的因果关系可能会被推翻。通常情况下，c引起e，但是c和d却不能引起e。例如，一个主体认为转动点火开关会使电动机启动，但是当该主体知道电池已经没电了，则不会相信转动点火开关会引起电动机启动。在人工智能中，这就是著名的条件问题(qualification problem)。由于一些潜在的相关因素通常是不确定的，所以进行明确的推理并不有效。因此，当进行因果推理时，人们通常会“跳”到结论，并在需要时推翻一些结论。类似地，从证据到原因的推理是非单调的。如果一个主体观察到一些结果e，则可以假设一个可能的原因c。由事实到原因的推理是溯因推理。对于某些事实来说，如果没有更好的解释，则接受溯因解释。然而，当产生一些新的解释时，旧的解释也可能会被丢弃。

6.2. 高效性和可解释性

从计算的角度看，单调性是经典逻辑的一个重要性质。它意味着利用知识的子集进行局部计算得到的每一个结论都等于利用所有知识进行全局计算得到的结论。然而，这一性质在非单调推理中并不成立，因此其计算效率可能非常低。在提高计算效率方面，比起其他一些非单调形式体系（如缺省逻辑和限制逻辑等），形式论辩已被证明是一个良好候选。其原因在于，在形式论辩中，可以采取分而治之的策略，以及依据论证图中节点的可达关系，最大限度地利用已有的计算结果[39]。在人工智能中，因果推理的另一个重要特性是可解释性。传统的非单调形式系统在用于解释方面并不理想，因为其中所有的证明都并非以人类可理解的方式来表达的。由于解释的目的是为了让人们能够更好地理解，比较和对比论证的认知过程具有十分重要的意义[37]。以辩护和论证对话的方式，论辩通过交换论证的方式提供了这样的一种途径[40]。

6.3. 与机器学习方法的联系

在可解释人工智能中，包含两个部分：可解释模型和解释接口。后者包括直接来自模型的自反解释和来自对用户信念进行推理的理性解释。为了实现这个目的，一种自然的方式是将论辩与机器学习结合起来。其中，知识通过机器学习获得，而推理和解释则通过论辩来实现。由于论辩提供了一种在不一致情景下进行各种推理的一般形式，并且可以与概率和模糊性等一些不确定性度量相结合，因此它能够灵活地对从数据中得到的知识进行建模。当机器学习一些特性并做出解释时，例如，“这张脸很生气，因为它与这些例子相似，而与另一些例子不同”，这就是一个论证，它可能会受到其他论证的攻击。并且，要衡量类似“愤怒”这样的词所描述的不确定性，人们可以选择使用可能性论辩或概率论辩[41]。不同的解释可能相互冲突。例如，在某些情况下，我们可能会采用支持某个选择的特定示例或故事，从而拒绝对另一基于分析、案例和数据的次优选择。通过使用论辩图，可以方便地对这类支持和攻击关系进行建模，从而计算不同选择下冲突论证的状态。

7. 复杂实验中的因果推断

因果推断的潜在结果框架始于一个假想的实验，在该实验中，实验者可以将每个样本分配到多个处理水平。每个样本都有与这些处理水平相对应的潜在结果, 而因果作用就是对同一组样本之间潜在结果的比较。这种方法有时被称为实验主义者的因果推断方法[42]。读者可进一步参考文献[43–46]。

7.1. 随机因子实验

Neyman [47]首先用数学严格讨论了如下的随机化模型。一个实验有n个样本，实验者随机分配$(n_1, \ldots, n_J)$个样本接受处理水平$(1, \ldots, J)$，其中$n = \sum_{j=1}^{J} n_j$。样本i有潜在结果$\{Y_i(1), \ldots, Y_i(J)\}$：如果样本$i$接收处理水平$j$，那么$Y_i(j)$是对应的结果。基于潜在结果，我们可以定义因果作用。比如，干预水平j和j'之间的比较为$\tau(j, j') = n^{-1}\sum_{i=1}^{n}\{Y_i(j) - Y_i(j')\}$。如果样本$i$实际接收到了处理水平$j$，则定义二值指标$T_i(j)$为1，用$Y_i = \sum_{j=1}^{J} T_i(j) Y_i(j)$表示样本$i$的观测结果。根据观测数据$\{T_i(1), \ldots, T_i(J)\}^{n}{}_{i=1}$，Neyman [47]建议使用$\hat{\tau}(j, j')=n_j^{-1}\sum_{i=1}^{n} T_i(j) Y_i - n_{j'}^{-1}\sum_{i=1}^{n} T_i(j') Y_i$作为$\tau(j, j')$的估计量。他证明$\hat{\tau}(j, j')$是无偏的，方差为$S^2(j)/n_j + S^2(j')/n_{j'} - S^2(j - j')/n$，其中$S^2(j)$、$S^2(j')$和$S^2(j - j')$是$Y_i(j)$、$Y_i(j')$和$Y_i(j) - Y_i(j')$的样本方差。注意，所有潜在结果都是固定的，这个问题的随机性来自于二值的处理指标。Neyman [47]进一步讨论了方差估计和大样本置信区间等问题。

我们可以将Neyman [47]的框架推广到一个更广泛的因果作用$\tau = n^{-1}\sum_{i=1}^{n}\tau_i$，其中$\tau_i = \sum_{j=1}^{J} c_j Y_i(j)$表示个体作用，

而c_j是满足$\sum_{j=1}^{J} c_j = 0$的比较矩阵。只要适当选择比较矩阵，这个定义就包含了方差分析[48]和因子实验[49,50]。此外，只要适当选择样本子集，这个定义也包含了亚组分析、事后分层[51]和同侪效应[52]。文献[53]提供了在这种框架下渐近统计推断所需的中心极限定理的一般形式。文献[54]讨论了裂区设计，文献[55]讨论了更广的实验设计。

7.2. 协变量在分析实验数据中的作用

Neyman [47]的随机化模型也允许在没有强建模假设时使用协变量提高估计精度。在处理为二值的情况下，对于样本i，用$\{Y_i(1), Y_i(0)\}$表示潜在结果，T_i表示二值处理变量，x_i表示协变量。平均因果作用$\tau = n^{-1}\sum_{i=1}^{n}\{Y_i(1), Y_i(0)\}$的一个无偏估计量为$\hat{\tau} = n_1^{-1}\sum_{i=1}^{n} T_i Y_i - n_0^{-1}\sum_{i=1}^{n}(1-T_i)Y_i$。Fisher [56]建议使用协方差分析以提高估计精度；也就是用Y_i对T_i和x_i拟合一个最小二乘，然后使用T_i的系数去估计τ。在文献[47]的模型下，文献[57]证明Fisher的协方差分析的估计并不一定好，它的估计精度甚至比$\hat{\tau}$还低，而且最小二乘法可能给出不相合的方差估计。文献[58]提出一个简单的修正：第一步，中心化协变量，使平均值为零，$\bar{x} = 0$；第二步，用Y_i对$(T_i, x_i, T_i \times x_i)$拟合一个最小二乘，然后使用$T_i$的系数去估计$\tau$；第三步,使用Eicker-Huber-White方差估计值[59–61]。在大样本下，文献[58]的估计至少和$\hat{\tau}$一样有效，并且Eicker-Huber-White方差估计是$\hat{\tau}$真实方差的保守估计。

文献[62]将分析推广到高维协变量，并用文献[63]提出的最小绝对收缩和选择算子（least absolute shrinkage and selection operator, ASSO）替换了最小二乘。文献[64]研究了文献[58]中最小二乘估计值的理论边界，考虑了协变量个数可能发散的情形。文献[65]用Y_i在$(T_i, x_i, T_i \times x_i)$上的最小二乘拟合来研究处理作用的异质性。文献[66]讨论了因子实验中的协变量调整，而文献[67]讨论了更广的实验设计中的协变量调整。

7.3. 协变量在实验设计中的作用

分析人员可以使用协变量提高估计效率。与之对偶，设计人员可以用协变量改善协变量平衡，从而提高估计效率。文献[68]暗示了重新随机化的思想，也就是只接受那些能确保协变量平衡的随机分配。考虑一个特殊的例子：我们接受随机分配当且仅当$\hat{\tau}'_x\{nS_x^2/(n_1 n_0)\}^{-1}\hat{\tau}_x \leqslant a$，其中$\hat{\tau}_x = n_1^{-1}\sum_{i=1}^{n} T_i x_i - n_0^{-1}\sum_{i=1}^{n}(1-T_i)x_i$, $S_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$，且$a > 0$，并且是一个预定常数时。文献[69]正式地讨论了这个重新随机化在处理组和对照组有相同样本量，协变量为正态分布，且处理作用为常数时的统计性质。文献[70]在没有这些假设的情况下研究了$\hat{\tau}$的渐近理论。文献[70]证明$\hat{\tau}$具有一个非正态的极限分布，并且它在重新随机化下比在完全随机化下更接近于τ。文献[70]中的结果表明当$a \approx 0$时，重新随机化下$\hat{\tau}$的渐近方差与完全随机化下[58]提出的估计量的渐近方差几乎一样。因此,我们可以把重新随机化看作是回归调整的对偶。

文献[71]提出一个能反映不同协变量重要性的重新随机化方案,文献[70]分析了该方案的渐近性质。文献[72,73]将重新随机化扩展到因子实验,文献[74]提出了序贯重新随机化。

7.4. 结语

受到文献[47]的启发，这一节重点回顾了随机实验中估计量的重复采样性质。另外，在所有的样本都满足$Y_i(1) = \ldots = Y_i(J)$的强零假设下，对于任何检验统计量和任何实验设计，Fisher随机化检验都是在有限样本下精确的假设检验[46,75,76]。文献[77,78]提出了在随机化检验中使用协变量调整的方法,文献[69]提出用随机化检验分析重新随机化。文献[79–81]将随机化检验应用于有干扰的实验。文献[48,50,82]讨论了随机化检验在弱零假设下的性质。文献[83–85]提出通过反转一系列随机化检验，以构建准确的置信空间。文献[86]从缺失数据的角度讨论了不同的统计推断框架。

8. 观察性研究中的工具变量和阴性对照方法

在很多科学研究中，人们最终的目的是评价一种处理或者暴露因素对一种结果或者响应变量的因果作用。自文献[75]提出随机化试验以来，其成为一种非常有效和广泛使用的评价因果作用的方法。但是，在很多研究中，由于伦理、经济或者不依从因素的制约，随机化试验并不适用或者代价太高；反而，在这样的研究中，观察性研究提供了更重要的数据来源和研究方法。不过，在观察性研究中，推断因果作用仍然面临很多挑战。最常见的问题是有混杂因素存在。混杂因素是指同时影响关心的处理和结果的因素或协变量。如果混杂因素被观测到，可以使用标准的统计推断方法做调整。然而，如果有混杂因素未被观测到，这些标准的调整方法通常失效，甚至会导致悖论。著名的Yule-Simpson悖论[19,20]就是一个例子。关于混杂的概念，文献[87,88]提供了很好的文献回顾。关于已观测混杂因素的调整，文献[2,89,90]讨论了常用的调整方法，例如，倾

向得分和逆概率加权，回归调整和双稳健方法。本文着重回顾两种针对未观测混杂因素的调整方法，一种是经典的工具变量方法；另一种是近期引起人们重视的阴性对照方法。本文使用X和Y分别表示关心的处理和结果变量，U是未观测的混杂变量。为了符号上的方便，忽略已观测的混杂变量，当然，本文的结果在条件为已观测的混杂变量时仍然成立。使用小写的英文字母表示相应变量的一个观测值，例如，y表示Y的一个观测值。

工具变量方法在1928年由文献[91,92]提出，现在，该方法已成为经济学、社会学、流行病和生物医学等学科中的观察性研究的重要方法。除了关心的处理和结果变量，这种方法还需要额外观测一个工具变量，用Z表示，其需要满足如下条件：

（1）Z对Y没有作用：$Z \perp Y \mid (X, U)$（无直接作用）；

（2）Z和U独立：$Z \perp U$（独立性）；

（3）Z和X相关：$Z \perp X$（相关性）。

即使在这三个假定下，也只能得到X对Y的因果作用的上下界，而不能唯一确定因果作用，也就是不能识别[93,94]。为了识别因果作用，需要额外的信息或者模型假定。结构方程模型[91,95]和结构均值模型[96]是常用的模型。这两个模型通过假定因果作用的同质性，即在不同个体上的因果作用是常数，从而得到识别性[97]。一个例子是常见的线性模型$(Y|X, U) = \alpha + \beta X + U$，实际上假定了在所有个体上的因果作用都是$\beta$；根据这个模型可以得到人们熟知的工具变量识别公式$\beta_{\text{iv}} = \sigma_{sy}/\sigma_{xs}$。除了同质性的模型假定，在一些实际问题中可以假定单调性。例如，在一些临床试验中，Z表示处理分配，X表示个体实际接受或者采取的处理，由于有不依从存在，X可能和Z不完全相同，但是有时可以合理地假设Z对X的作用是单调的：$X_{z=1} \geqslant X_{z=0}$，即没有个体会采取与其被分配的相反的处理。这里，$X_Z$表示在处理分配$Z = z$下潜在接受的处理。在单调性假定下，可以识别依从组的因果作用：$(\text{CACE}) = E(Y_1 - Y_0|X_1 = 1, X_0 = 0)$ [98]。此外，在一些应用场景，比如统计遗传学中，当有多个或高维工具变量时，如何选择工具变量变得很重要[99,100]。

在实际研究中，寻找一个满足如上三条工具变量假定的变量并不容易，而且工具变量方法对这三条假定都非常敏感。如何验证工具变量假定也是一个重要话题，例如，使用工具变量不等式做检验，见文献[94,101]。如果工具变量假定不成立，因果作用通常不可识别，工具变量方法也有偏，在这种情况下，求因果作用的界和敏感性分析方法更加稳健[102,103]。

由文献[104–106]提出和建立的基于阴性对照的因果推断方法提供了一个新的观察性研究的工具，也提供了一个补救工具变量可能失效的方法。阴性对照变量分为两类：阴性对照结果和阴性对照暴露，分别用W和Z表示，它们分别是一个辅助的结果变量和暴露变量，分别需要满足如下的条件：

$W \perp\!\!\!\perp X|U,\ W \not\perp\!\!\!\perp U,\ Z \perp\!\!\!\perp Y|(U, X),\ Z \perp\!\!\!\perp W|(U, X).$

阴性对照暴露Z可以看作工具变量的推广，其满足对Y无直接作用的假定，但不必和未观测的混杂U独立，这一点不同于工具变量。给定一对阴性对照暴露和结果变量，在一定的正则性条件下，文献[104,106]证明了平均因果作用的非参数识别性，即不需要假定参数化的模型。这里，作为示例，考虑线性模型$E(Y \mid X, U) = \alpha + \beta X + U$，并假定$E(W|U)$也是关于$U$的线性模型，那么，$\beta$的阴性对照识别结果为：

$$\beta_{\text{nc}} = \frac{\sigma_{xw}\sigma_{zy} - \sigma_{xy}\sigma_{zw}}{\sigma_{xw}\sigma_{xz} - \sigma_{xx}\sigma_{zw}}$$

这个识别公式也适用于是工具变量的情况，因为，如果Z是一个工具变量，那么$Z \perp U$且$\sigma_{zw} = 0$。因此，工具变量方法可以看作阴性对照方法的特例，当Z不满足工具变量的假定时，使用阴性对照结果W来消除Z造成的偏差。但是，阴性对照比工具变量需要的假定更弱，在实际中例子很多。例如，文献[107,108]回顾了观察性研究中的许多阴性对照的实例；文献[105,109]还指出，在很多时间序列研究中，阴性对照成立。例如，在空气污染研究中，当前时间的空气污染不影响前一段时间的公共卫生状况，而后一段时间的空气污染也不影响当前时间的公共卫生状况，即公共卫生状况对空气污染没有反馈作用。在这种情况下，前一段时间的公共卫生状况和后一段时间的空气污染可以分别作为阴性对照结果和阴性对照暴露变量。

文献[104–106]提出的阴性对照方法需要两个阴性对照变量，当只有一个阴性对照结果或者一个阴性对照暴露时，因果作用不可识别。

在这种情况下，文献[107,109,110]研究了使用单个阴性对照检验混杂是否存在和减小偏差的方法，但不能识别因果作用。当阴性对照变量很多时，文献[111,112]通过因子分析的方法消除偏差，但需要依赖严苛的参数化模型假定。

总之，观察性研究中如何处理混杂因素仍然是一个难题。尽管使用工具变量和阴性对照等辅助变量方法可以大幅提高因果作用的识别性，但是工具变量和阴性对照的假

定不能通过观察数据检验，而需要先验知识或者额外的研究确认。把先验知识和观察数据结合起来推断因果作用是观察性研究的一个可行的方向。现代大数据为因果推断提供了非常广阔的研究场景，但是，大数据通常都是观察性数据，而不是试验数据。因此，在大数据研究中处理混杂因素很重要。如何把先验知识和大数据结合起来更有效地推断因果作用需要更深入的研究。

9. 有干扰下的因果推断

个体处理稳定性假设是传统的潜在结果模型中的一个重要假设，它假设个体之间是没有干扰的[76]。但是，在很多试验和观测性研究中，个体之间会相互影响，从而造成了个体间的干扰。例如，在教育学或者社会学研究中，参与培训计划的学生会通过课堂外的交流影响没有参与的学生[113,114]。在流行病学中，传染病的预防措施会使得人们被传染的概率降低，因此即便没有接受预防的人群也会受到影响[115,116]。在这些研究中，个体接受到的处理不仅会对自己的结果变量有直接作用，还会对其他个体的结果变量有溢出作用。在实际问题中，直接作用和溢出作用有着重要的科学意义和社会意义，它们能帮助我们更好地理解因果作用的机制，从而能对政策或者方案的实施有指导作用。

在干扰存在的时候，一个个体的潜在结果的数量会随着整个样本中个体的数量呈指数增长。因此，在对干扰的结构没有约束时，我们无法得到直接和溢出作用的估计。有许多文献对干扰下的因果作用的估计进行了研究[117]，其中一个重要的方向是将干扰限制在一些不重叠的小群体中，假设不同群体之间的个体没有干扰[52,114,118–122]。这种假定被称为部分干扰假定[114]。近来，许多研究者试着减弱这个假定去处理更一般的干扰结构[123–126]。在有干扰的情况下，因果作用的方差估计会变得更加困难。文献[118]指出，即便在部分干扰假定下，直接和溢出作用的估计依然是很困难的。在不加任何模型假设的条件下，一个能得到有效的方差估计的方法是假设个体的结果变量只依赖于自己接受到的处理和其他个体接收到的处理的某个函数。例如，个体的结果只依赖于自己是否接受了处理，以及有多少个其他个体接收到了处理。

另外一个方向的研究侧重于如何根据干扰的结构去设计试验估计因果作用。在部分干扰的假定下，文献[118]提出两阶段随机化试验去估计直接和溢出作用。在更复杂的干扰结构下，研究者提出了各种各样的试验设计去得到因果作用的点估计和方差估计[127–129]。

对于干扰情况下的统计推断，文献[130,131]依赖于关于潜在结果的模型，文献[79]对没有溢出作用的零假设提出了一个条件随机化检验的方法。文献[80]将这个方法推广到了一大类关于焦点个体的假设检验中。基于这些方法，文献[132]提出了得到有效的条件检验的一般程序。

虽然研究者对个体间的干扰提出了许多方法，但是这方面的研究依然存在着很多挑战。首先，个体干扰下的极限性质还不够完善。文献[133]研究了均值差估计量在有限制的干扰假定下的一致性。文献[134]在部分干扰假定和分层干扰的假定下得到了直接作用和溢出作用的中心极限定理。但是在一般的干扰结构下，即便是最简单的均值差估计量，极限性质都是不清楚的。其次，在有复杂数据时，个体间的干扰会变得更加难以处理。文献[120,121,135,136]考虑了干扰情况下的不依从问题，文献[137]研究了干扰情况下生存时间数据的分析方法。但是，对于其他类型的复杂数据，比如缺失数据和测量误差还没有方法来进行处理。最后，绝大部分的文献关心的是直接作用和溢出作用，但是个体间的干扰在其他的问题中也会存在，比如中介分析（见文献[138]）和纵向数据分析。在这些问题中，我们关心的因果作用是不一样的。因此，我们需要将这些问题中的一些常用方法进行推广去处理个体间的干扰。

Compliance with ethics guidelines

Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences. New York: Cambridge University Press; 2015.

[2] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61(4):962–73.

[3] Kuang K, Cui P, Li B, Jiang M, Yang S, Wang F. Treatment effect estimation with data-driven variable decomposition. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence; 2017 Feb 4–9; San Francisco, CA, USA, 2017.

[4] Athey S, Imbens GW, Wager S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. J R Stat Soc Ser B (Stat Methodol) 2018;80(4):597–623.

[5] Kuang K, Cui P, Li B, Jiang M, Yang S. Estimating treatment effect in the wild via differentiated confounder balancing. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017 Aug 13–17; Halifax, NS, Canada. p. 265–74.

[6] Imai K, Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. J Am Stat Assoc 2004;99(467):854–66.

[7] Egami N, Imai K. Causal interaction in factorial experiments: application to conjoint analysis. J Am Stat Assoc 2019;114(526):529–40.

[8] Louizos C, Shalit U, Mooij JM, Sontag D, Zemel R, Welling M. Causal effect inference with deep latent-variable models. In: Proceedings of Advances in Neural Information Processing Systems 30; 2017 Dec 4–9; Long Beach, CA, USA. p. 6446–56.

[9] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika 2009;96(1):187–99.

[10] Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. Am J Epidemiol 2019;188(1):250–7.

[11] Kuang K, Cui P, Athey S, Xiong R, Li B. Stable prediction across unknown environments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018 Aug 19–23; London, UK. p. 1617–26.

[12] Zhuang Y, Wu F, Chen C, Pan Y. Challenges and opportunities from big data to Knowledge in AI2.0. Front Inf Technol Elec Eng 2017;18(1):3–14.

[13] Pan Y. 2018 special issue on artificial intelligence 2.0: theories and applications. Front Inf Technol Elec Eng 2018;19(1).

[14] Hoerl C, McCormack T, Beck SR, editors. Understanding counterfactuals, understanding causation: issues in philosophy and psychology. New York: Oxford University Press; 2011.

[15] Pearl J, Glymour M, Jewell NP. Causal inference in statistics: a primer. Hoboken: John Wiley & Sons; 2016.

[16] Daniel RM, De Stavola BL, Vansteelandt S. Commentary: the formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? Int J Epidemiol 2016;45(6):1817–29.

[17] Pearl J. Causal and counterfactual inference. Forthcoming section in the handbook of rationality. Cambridge: MIT press; 2018.

[18] Goldfeld K. Considering sensitivity to unmeasured confounding: part 1 [Internet]. New York: Keith Golgfeld; 2019. Jan 2 [cited 2019 Jun 1]. Available from:https://www.rdatagen.net/post/what-does-it-mean-if-findings-aresensitive- to-unmeasured-confounding/.

[19] Yule GU. Notes on the theory of association of attributes in statistics. Biometrika 1903;2(2):121–34.

[20] Simpson EH. The interpretation of interaction in contingency tables. J R Stat Soc B 1951;13(2):238–41.

[21] Chen H, Geng Z, Jia J. Criteria for surrogate end points. J R Stat Soc Series B Stat Methodol 2007;69(5):919–32.

[22] Geng Z, Liu Y, Liu C, Miao W. Evaluation of causal effects and local structure learning of causal networks. Annu Rev Stat Appl 2019;6(1):103–24.

[23] Pearl J. Is scientific knowledge useful for policy analysis? A peculiar theorem says: no. J Causal Infer 2014;2(1):109–12.

[24] Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125(7):605–13.

[25] Xu L, Pearl J. Structuring causal tree models with continuous variables. in Proc. 3rd Annu. Conf. Uncertainty in Artificial Intelligence, Seattle, USA, pp. 170-179, 1987.

[26] Xu L. Deep bidirectional intelligence: alphazero, deep IA-search, deep IAinfer, and TPC causal learning. Appl Inf 2018;5(5).

[27] Xu L. Machine learning and causal analyses for modeling financial and economic data. Appl Inf 2018;5(11).

[28] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd ed. Cambridge: MIT Press; 2001.

[29] Pearl J. Causality: models, reasoning, and inference. Cambridge: Cambridge University Press; 2000.

[30] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Appl Inform 2016;3(1):3.

[31] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7:2003–30.

[32] Zhang K, Hyvärinen A. On the identifiability of the post-nonlinear causal model. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; 2009 Jun 18–21; Montreal, QC, Canada. Arlington: AUAI Press; 2019. p. 647–55.

[33] Hoyer PO, Janzing D, Mooij JM, Peters J, Scholkopf B. Nonlinear causal discovery with additive noise models. In: Proceedings of International Conference on Neural Information Processing Systems; 2008 Dec 8–13; Vancouver, BC, Canada. p. 689–96.

[34] Zhang K, Hyvärinen A. Causality discovery with additive disturbances: an information-theoretical perspective. In: Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J, editors. Machine learning and knowledge discovery in databases. Berlin: Springer; 2009. p. 570–85.

[35] Zhang K, Schölkopf B, Muandet K, Wang Z. Domain adaptation under target and conditional shift. In: Proceedings of the 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA. p. 819–27.

[36] Baroni P, Gabbay DM, Giacomin M, Van der Torre L. Handbook of formal argumentation. London: College Publications; 2018.

[37] Osborne J. Arguing to learn in science: the role of collaborative, critical discourse. Science 2010;328(5977):463–6.

[38] Shoham Y. Nonmonotonic reasoning and causation. Cogn Sci 1990;14 (2):213–52.

[39] Liao B, Jin L, Koons RC. Dynamics of argumentation systems: a division-based method. Artif Intell 2011;175(11):1790–814.

[40] Sklar EI, Azhar MQ. Explanation through argumentation. In: Proceedings of the 6th International Conference on Human-Agent Interaction; 2018 Dec 15– 18; Southampton, UK. p. 277–85.

[41] Fazzinga B, Flesca S, Furfaro F. Complexity of fundamental problems in probabilistic abstract argumentation: beyond independence. Artif Intell 2019;268:1–29.

[42] Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence; 2010 Jul 8–11; Catalina Island, CA, USA. p. 425–32.

[43] Kempthorne O. The design and analysis of experiments. New York: Wiley; 1952.

[44] Scheffe H. The analysis of variance. New York: John Wiley & Sons; 1959.

[45] Hinkelmann K, Kempthorne O. Design and analysis of experiments: volume 1: introduction to experimental design. 2nd ed. New York: John Wiley & Sons; 2007.

[46] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.

[47] Splawa-Neyman J. On the application of probability theory to agricultural experiments: essay on principles. Section 9. Stat Sci 1990;5(4):465–72.

[48] Ding P, Dasgupta T. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika 2018;105(1):45–56.

[49] Dasgupta T, Pillai NS, Rubin DB. Causal inference from 2?? factorial designs by using potential outcomes. J R Stat Soc Series B Stat Methodol 2015;77 (4):727–53.

[50] Wu J, Ding P. Randomization tests for weak null hypotheses. 2018. arXiv:1809.07419.

[51] Miratrix LW, Sekhon JS, Yu B. Adjusting treatment effect estimates by poststratification in randomized experiments. J R Stat Soc Series B Stat Methodol 2013;75(2):369–96.

[52] Li X, Ding P, Lin Q, Yang D, Liu JS. Randomization inference for peer effects. J Am Stat Assoc 2019:1–31.

[53] Li X, Ding P. General forms of finite population central limit theorems with applications to causal inference. J Am Stat Assoc 2017;112(520):1759–69.

[54] Zhao A, Ding P, Mukerjee R, Dasgupta T. Randomization-based causal inference from split-plot designs. Ann Stat 2018;46(5):1876–903.

[55] Mukerjee R, Dasgupta T, Rubin DB. Using standard tools from finite population sampling to improve causal inference for complex experiments. J Am Stat Assoc 2018;113(522):868–81.

[56] Fisher R. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.

[57] Freedman DA. On regression adjustments to experimental data. Adv Appl Math 2008;40(2):180–93.

[58] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. Ann Appl Stat 2013;7(1):295–318.

[59] Eicker F. Limit theorems for regressions with unequal and dependent errors. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967 Jun 21–Jul 18; Berkeley, CA, USA. Berkeley: University of California Press; 1967. p. 59–82.

[60] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967 Jun 21–Jul 18; Berkeley, CA, USA. p. 221–33.

[61] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 1980;48(4):817–38.

[62] Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Natl Acad Sci USA 2016;113(27):7383–90.

[63] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267–88.

[64] Lei L, Ding P. Regression adjustment in completely randomized experiments with a diverging number of covariates. 2018. arXiv:1806.07585.

[65] Ding P, Feller A, Miratrix L. Decomposing treatment effect variation. J Am Stat Assoc 2019;114(525):304–17.

[66] Lu J. Covariate adjustment in randomization-based causal inference for 2?? factorial designs. Stat Probab Lett 2016;119:11–20.

[67] Middleton JA. A unified theory of regression adjustment for design-based inference. 2018. arXiv:1803.06011.

[68] Cox DR. Randomization and concomitant variables in the design of experiments. In: Anderson TW, Styan GHP, Kallianpur GG, Krishnaiah PR, Ghosh JK, editors. Statistics and probability: essays in honor of CR Rao. Amsterdam: North-Holland; 1982. p. 197–202.

[69] Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. Ann Stat 2012;40(2):1263–82.

[70] Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment-control experiments. Proc Natl Acad Sci USA 2018;115 (37):9157–62.

[71] Morgan KL, Rubin DB. Rerandomization to balance tiers of covariates. J Am Stat Assoc 2015;110(512):1412–21.

[72] Branson Z, Dasgupta T, Rubin DB. Improving covariate balance in 2?? factorial designs via rerandomization with an application to a New York City department of education high school study. Ann Appl Stat 2016;10 (4):1958–76.

[73] Li X, Ding P, Rubin DB. Rerandomization in 2?? factorial experiments. 2018. arXiv:1812.10911.

[74] Zhou Q, Ernst PA, Morgan KL, Rubin DB, Zhang A. Sequential rerandomization. Biometrika 2018;105(3):745–52.

[75] Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd; 1935.

[76] Rubin DB. Comment on ''randomization analysis of experimental data: the

Fisher randomization test". J Am Stat Assoc 1980;75(371):591–3.

[77] Tukey JW. Tightening the clinical trial. Control Clin Trials 1993;14(4):266–85.

[78] Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. Stat Sci 2002;17(3):286–327.

[79] Aronow PM. A general method for detecting interference between units in randomized experiments. Sociol Methods Res 2012;41(1):3–16.

[80] Athey S, Eckles D, Imbens GW. Exact p-values for network interference. J Am Stat Assoc 2018;113(521):230–40.

[81] Basse G, Feller A, Toulis P. Exact tests for two-stage randomized designs in the presence of interference. 2017. arXiv:1709.08036.

[82] Ding P. A paradox from randomization-based causal inference. Stat Sci 2017;32(3):331–45.

[83] Rosenbaum PR. Exact confidence intervals for nonconstant effects by inverting the signed rank test. Am Stat 2003;57(2):132–8.

[84] Rigdon J, Hudgens MG. Randomization inference for treatment effects on a binary outcome. Stat Med 2015;34(6):924–35.

[85] Li X, Ding P. Exact confidence intervals for the average causal effect on a binary outcome. Stat Med 2016;35(6):957–60.

[86] Ding P, Li F. Causal inference: a missing data perspective. Stat Sci 2018;33 (2):214–37.

[87] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat. Sci 1999;14:29–46.

[88] Greenland S, Pearl J. Adjustments and their consequences—collapsibility analysis using graphical models. Int Stat Rev 2011;79(3):401–26.

[89] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55.

[90] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952;47(260):663–85.

[91] Wright PG. Tariff on animal and vegetable oils. New York: Macmillan; 1928.

[92] Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour 1997;32 (3):441–62.

[93] Manski CF. Nonparametric bounds on treatment effects. Am Econ Rev 1990;80(2):319–23.

[94] Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Am Stat Assoc 1997;92(439):1171–6.

[95] Goldberger AS. Structural equation methods in the social sciences. Econometrica 1972;40(6):979–1001.

[96] Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Method 1994;23 (8):2379–412.

[97] Hernán MA, Robins JM. Causal inference. Boca Raton: Chapman & Hall; 2011.

[98] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;91(434):444–55.

[99] Lin W, Feng R, Li H. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. J Am Stat Assoc 2015;110(509):270–88.

[100] Kang H, Zhang A, Cai TT, Small DS. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. J Am Stat Assoc 2016;111(513):132–44.

[101] Wang L, Robins JM, Richardson TS. On falsification of the binary instrumental variable model. Biometrika 2017;104(1):229–36.

[102] Manski CF, Pepper JV. Monotone instrumental variables: with an application to the returns to schooling. Econometrica 2000;68(4):997–1010.

[103] Small DS. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. J Am Stat Assoc 2007;102(479):1049–58.

[104] Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika 2018;105 (4):987–93.

[105] Miao W, Tchetgen Tchetgen E. Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. Am J Epidemiol 2017;185(10):950–3.

[106] Miao, W., Shi, X., Tchetgen E., Tchetgen. A. confounding cridge approach for couble negative control inference on causal effects 2018. arXiv:1808.04945.

[107] Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 2010;21(3):383–8.

[108] Smith GD. Negative control exposures in epidemiologic studies. Epidemiology 2012;23(2):350–1.

[109] Flanders WD, Strickland MJ, Klein M. A new method for partial correction of residual confounding in time-series and other observational studies. Am J Epidemiol 2017;185(10):941–9.

[110] Rosenbaum PR. The role of known effects in observational studies. Biometrics 1989;45(2):557–69.

[111] Wang J, Zhao Q, Hastie T, Owen AB. Confounder adjustment in multiple hypothesis testing. Ann Stat 2017;45(5):1863–94.

[112] Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics 2012;13(3):539–52.

[113] Hong G, Raudenbush SW. Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. J Am Stat Assoc 2006;101(475):901–10.

[114] Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J Am Stat Assoc 2006;101 (476):1398–407.

[115] Halloran ME, Struchiner CJ. Causal inference in infectious diseases. Epidemiology 1995;6(2):142–51.

[116] Halloran ME, Struchiner CJ. Study designs for dependent happenings. Epidemiology 1991;2(5):331–8.

[117] Halloran ME, Hudgens MG. Dependent happenings: a recent methodological review. Curr Epidemiol Rep 2016;3(4):297–305.

[118] Hudgens MG, Halloran ME. Toward causal inference with interference. J Am Stat Assoc 2008;103(482):832–42.

[119] Basse G, Feller A. Analyzing two-stage experiments in the presence of interference. J Am Stat Assoc 2018;113(521):41–55.

[120] Forastiere L, Mealli F, VanderWeele TJ. Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using bayesian principal stratification. J Am Stat Assoc 2016;111 (514):510–25.

[121] Kang H, Imbens G. Peer encouragement designs in causal inference with partial interference and identification of local average network effects. 2016. arXiv:1609.04464.

[122] Rigdon J, Hudgens MG. Exact confidence intervals in the presence of interference. Stat Probab Lett 2015;105:130–5.

[123] Aronow PM, Samii C, 2013. Estimating average causal effects under interference between units. arXiv preprint arXiv:1305.61563, 16.

[124] Aronow PM, Samii C. Estimating average causal effects under general interference, with application to a social network experiment. Ann Appl Stat 2017;11(4):1912–47.

[125] Choi D. Estimation of monotone treatment effects in network experiments. J Am Stat Assoc 2017;112(519):1147–55.

[126] Forastiere L, Airoldi EM, Mealli F. Identification and estimation of treatment and interference effects in observational studies on networks. 2016. arXiv:1609.06245.

[127] Eckles D, Karrer B, Ugander J. Design and analysis of experiments in networks: reducing bias from interference. J Causal Inference 2017;5(1).

[128] Eckles D, Kizilcec RF, Bakshy E. Estimating peer effects in networks with peer encouragement designs. Proc Natl Acad Sci USA 2016;113(27):7316–22.

[129] Jagadeesan R, Pillai N, Volfovsky A. Designs for estimating the treatment effect in networks with interference. 2017. arXiv:1705.08524.

[130] Bowers J, Fredrickson MM, Panagopoulos C. Reasoning about interference between units: a general framework. Polit Anal 2013;21(1):97–124.

[131] Toulis P, Kao E. Estimation of causal peer influence effects. In: Proceedings of 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA. p. 1489–97.

[132] Basse GW, Feller A, Toulis P. Randomization tests of causal effects under interference. Biometrika 2019;106(2):487–94.

[133] Sävje F, Aronow PM, Hudgens MG. Average treatment effects in the presence of unknown interference. 2017. arXiv:1711.06399.

[134] Liu L, Hudgens MG. Large sample randomization inference of causal effects in the presence of interference. J Am Stat Assoc 2014;109(505):288–301.

[135] Imai K, Jiang Z, Malani A. Causal inference with interference and noncompliance in two-stage randomized experiments. Princeton: Technical report Princeton University; 2018.

[136] Kang H, Keele L. Spillover effects in cluster randomized trials with noncompliance. 2018. arXiv:1808.06418.

[137] Loh WW, Hudgens MG, Clemens JD, Ali M, Emch ME. Randomization inference with general interference and censoring. 2018. arXiv:1803.02302.

[138] Vanderweele TJ, Hong G, Jones SM, Brown JL. Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention. J Am Stat Assoc 2013;108(502):469–82.