A Survey of Learning Causality with Data: Problems and Methods

RUOCHENG GUO, Computer Science and Engineering, Arizona State University
LU CHENG, Computer Science and Engineering, Arizona State University
JUNDONG LI, Computer Science and Engineering, Arizona State University
P. RICHARD HAHN, Department of Mathematics and Statistics, Arizona State University
HUAN LIU, Computer Science and Engineering, Arizona State University

The era of big data provides researchers with convenient access to copious data. However, we often have little knowledge of such data. The increasing prevalence of massive data challenges the traditional methods of learning causality because they were developed for the cases with limited amount of data and strong prior causal knowledge. This survey aims to close the gap between big data and learning causality with a comprehensive and structured review of both traditional and frontier methods followed by a discussion about some open problems of learning causality. We begin with preliminaries of learning causality. Then we categorize and revisit methods of learning causality for typical problems and different data types. After that, we discuss the connections between learning causality and machine learning. At the end, some open problems are presented to show the great potential of learning causality with data.

ACM Reference Format:
Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2010. A Survey of Learning Causality with Data: Problems and Methods. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 36 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Causality is a generic relationship between an effect and the cause that gives rise to it. It is hard to define, and we often only know intuitively about causes and effects. Because it rained, the streets were wet. Because the student did not study, he did poorly on the exam. Because the oven was hot, the cheese melted on the pizza. When it comes to learning causality with data, we need to be aware of the differences between statistical associations and causations. For example, when the temperatures are hot, the owner of an ice cream shop may observe high electric bills and also high sales. Accordingly, she would observe a strong association between the electric bill and the sales figures, but the electric bill was not *causing* the high sales — leaving the lights on in the shop over night would have no impact on sales. In this case, the outside temperature is the common cause of both the high electric bill and the high sales numbers, and we say that it is a *confounder* of the causality of the electricity usage on the ice cream sales.

The ability to learn causality is considered as a significant component of human-level intelligence and can serve as the foundation of AI [105]. Historically, learning causality has been studied

Authors' addresses: Ruocheng Guo, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, rguo12@asu.edu; Lu Cheng, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, lcheng35@asu.edu; Jundong Li, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, jundongl@asu.edu; P. Richard Hahn, Department of Mathematics and Statistics, Arizona State University, Tempe, AZ, 85281, prhahn@asu.edu; Huan Liu, Computer Science and Engineering, Arizona State University, Tempe, AZ, 85281, huan.liu@asu.edu.

1559-1131/2010/3-ART39 $15.00
https://doi.org/0000001.0000001

in myriad of high-impact domains including education [36, 60, 62, 80], medical science [94, 100], economics [71], epidemiology [61, 92, 117], meteorology [37], and environmental health [86]. Limited by the amount of data, solid prior causal knowledge was necessary for learning causality. Researchers performed studies on data collected through carefully designed experiments where solid prior causal knowledge is of vital importance [60]. Taking the *randomized controlled trials* as a prototype example [33], to study the efficacy of a drug, a patient would be randomly assigned to take the drug or not, which would guarantee that — *on average* — the treated and the un-treated (control) group are equivalent in all relevant respects, while ruling out the influence of any other factors. Then, the impact of the drug on some health outcome — say, the duration of a migraine headache — can be measured by comparing the average outcome of the two groups.

The purpose of this survey is to consider what new possibilities and challenges arise for learning about causality in the era of big data. As an example, consider that the possibility of unmeasured confounders — might be mitigated now as a vast amount of features can be measured. So, on one hand, it becomes possible for researchers to answer interesting causal questions with the help of big data. For instance, do positive Yelp[1] reviews drive customers to restaurants, or do they merely reflect popularity but not influence it? This causal question can be addressed by data from an extensive database maintained by Yelp. On the other hand, answering causal questions with big data leads to some unique new problems. For example, though public databases or data collected through web crawling or application program interfaces (APIs) are unprecedentedly large, we have very little intuition about what types of bias the dataset can suffer from — the data is more plentiful, but also perhaps more mysterious and, therefore, harder to model responsibly. At the same time, causal investigation is made more challenging by the same fundamental statistical difficulties that big data poses for other learning tasks (e.g., prediction). Perhaps the most notable example is the high-dimensionality of modern data [84], such as text data [70].

Many begin to investigate this intersection between big data and causal inquiry. Notable examples include but are not limited to [10, 38, 89, 96, 111]. In this survey, we aim instead to catalogue the different types of data that are available in this era and to provide an overview of the existing methods that attempt to answer causal questions using those data. As part of this effort, we will review the two primary formal frameworks for studying causality as well as the basic methodologies for learning causality, that underlie more advanced techniques designed for big data.

1.1 Overview and Organization

Broadly, machine learning tasks are either *predictive* or *descriptive* in nature. But beyond that we may want to understand something *causal*, imaging that we were able to modify some variables and rerun the data-generating process. These types of questions can also take two (related) forms: 1) How much would some variables (features or the label) change if we manipulate the value of another variable? and 2) By modifying the value of which variables could we change the value of another variable? These questions are referred to as *causal inference* questions and *causal discovery* questions, respectively [46, 110]. For learning causal effects, we investigate to what extent manipulating the value of a potential cause would influence a possible effect. Following the literature, we call the variable to be manipulated as *treatment* and the variable for which we observe the response as *the outcome*, respectively. One typical example is that *how much do hot temperatures raise ice cream sales*. For learning causal relationships, researchers attempt to determining whether there exists a causal relationship between a variable and another. In our temperature and ice cream example, it is clear that ice cream sales do not cause high temperatures, but in other examples it

[1] https://www.yelp.com/

may not be clear. For example, we may be interested in investigating the question like *whether a genetic disposition towards cancer should be responsible for individuals taking up smoking?*

In this survey, we aim to start from the data perspective and provide a comprehensive review on how to learn causality from massive data. Below, we present an outline of the topics that are covered in this survey. First, in Section 2, we introduce the preliminaries of learning about causality from data, which we will shorten to *learning causality* to encompass either causal inference or causal discovery. We mainly focus on the two most important formal frameworks, namely the *structural causal models* [103] and the *potential outcome framework* [99, 121]. Next, in Section 3 and 4 we go over the most common methodologies for learning causality from data. Specifically, in these two sections, the methods are categorized by the types of data they can handle. Section 3 focuses on the methods that are developed for the problem of learning causal effects (causal inference). Based on different types of data, these methods fall into three categories: methods for i.i.d data, non-i.i.d. data with the back-door criterion, and the data without it. Then, in Section 4, the widely used methods for learning causal relationships are discussed. According to the data type, we first cover the methods for discovering causal relationships between variables in i.i.d. data. Then, we describe the methods that can tackle the inter-dependencies in time series data. Afterwards, in Section 5, we aim to provide an aspect of how some previous work narrowed the gap between learning causality and machine learning. Specifically, we go over how the research in three subareas of machine learning, namely supervised and semi-supervised learning, domain adaptation and reinforcement learning can be connected to learning causality.

1.2 Data for Learning Causality

In this subsection, we discuss data and methods that are used for learning causal effects and relationships[2]. We start with the data types and methods for learning causal effects and then cover those for learning causal relationships.

Data for Learning Causal Effects. Here, we provide an overview of the types of data for learning causality, the problems that can be studied if the data is given, and the methods that can provide practical solutions. We introduce three types of data that can be applied to study learning causal effect. First, a standard dataset for learning causal effects $(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{y})$ includes a matrix of features $\boldsymbol{X}$ which is considered to provide enough information about the instances (satisfies the back-door criterion, see Section 2), a vector of treatments $\boldsymbol{d}$ and outcomes $\boldsymbol{y}$. With such a representation, this type of data is similar to what is often used for supervised learning. The only difference is that we are particularly interested in the causal effect of one feature D on the label or another feature as the outcome Y. For the second type, in addition to those that are present in the first type, there is auxiliary information about inter-dependence or interference between units such as links or temporal inter-dependencies between different data units (samples), represented by a matrix $\boldsymbol{A}$. Some special cases of this type of data can be attributed networks [85, 148], time series [39], and marked temporal point process [55]. Moreover, when there are unobserved confounders in the third type, we need the help of special causal variables, including the *instrumental variable (IV)*, the *mediator*, and the *running variable*. These special variables are defined by typical causal knowledge, thus specific methods can be applied for learning causal effect for such types of data (see Section 3).

Data for Learning Causal Relationships. We also describe two types of data for the study of (learning causal relationships) causal discovery. The first type is the multivariate data $\boldsymbol{X}$ along with a ground truth causal graph G for evaluation, with which we learn the causal graph. A special case is the bivariate data and the task reduces to distinguishing the cause from the effect [96]. The causal

[2]The data index and algorithm index for learning causality are described in Appendix

Problems	Data	Example Datasets	Methods
Learning causal effects	Datasets with, features, treatment and outcome $(\boldsymbol{X}, \boldsymbol{t}, \boldsymbol{y})$.	IHDP, Twins[3], Jobs[4]	Regression adjustment, Propensity score, Covariate balancing, Machine learning based methods,
	Datasets with features, treatment, outcome and special variable(s): $(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{y}, \boldsymbol{z})$	1980 Census Extract, CPS Extract[5]	Instrumental variable, Front-door criterion, RDD, Machine learning based
Learning causal relationships	Multivariate data with causal relationships, denoted by $\boldsymbol{X}$ with a causal graph G, including bivariate data with causal direction.	Abscisic Acid Signaling Network[6], Weblogs[7], SIDO [8]	Constraint-based, Score-based methods, Algorithms for FCMs.
	Multivariate time series $\{[x_{,1}(l), ..., x_{,J}(l)]\}_{l=1}^{L}$ with a causal graph G	PROMO[9]	

Table 1. Overview of this work in terms of the problems, data and methods.

graph is often defined by prior knowledge and could be incomplete. The second type of data for learning causal relationships is the multivariate time series data which also comes with a ground truth causal graph. We learn causal relationships between different variables [49].

1.3 Previous Work and Contributions

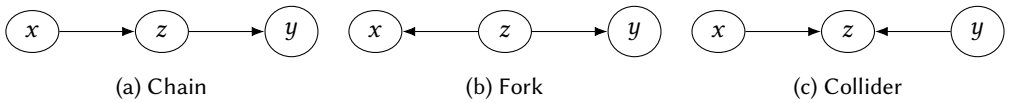
There are a number of other comprehensive surveys in the area of causal learning. Pearl [103] aimed to convey the fundamental theory of causality based on the structural causal models. Gelman [46] provided high-level opinions about the existing formal frameworks and problems for causal learning. Mooji et al. [96] focused on learning causal relationships for bivariate data. Spirtes and Zhang [133] summarized methods for learning causal relationships on both i.i.d. and time series data but they focus on several semi-parametric score based methods. Athey and Imbens [12] described decision trees and ensemble machine learning models for learning causal effects.

Different from previous work, this survey is structured around various data types, and what sorts of causal questions can be addressed with them. Specifically, we describe what types of data can be used for the study of causality, what are the problems that can be solved for each type of data and how they can be solved. In doing so, we aim to provide a bridge between the areas of machine learning, data mining, and causal learning in terms of terminologies, data, problems and methods.

1.4 Running Example

We consider a study of how Yelp ratings influence potential restaurant customers [6]. Yelp is a website where customers can share their reviews of a certain goods and services. Each review

[3]https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/IHDP
[4]http://users.nber.org/ rdehejia/data/nswdata2.html
[5]https://economics.mit.edu/faculty/angrist/data1/data/angkru95
[6]https://archive.ics.uci.edu/ml/datasets/Abscisic+Acid+Signaling+Network
[7]http://www.causality.inf.ethz.ch/repository.php?id=13
[8]http://www.causality.inf.ethz.ch/data/SIDO.html
[9]http://clopinet.com/causality/data/promo/



(a) Chain (b) Fork (c) Collider

Fig. 1. Three typical DAGs for conditional independence

includes an integer rating from 1 to 5 stars. For our purposes, the Yelp rating is our *treatment* variable and the number of customers (in some well-defined period) is the *outcome* variable. For simplicity, we consider these variables are binary. A restaurant receives an active treatment $t = 1$ if its rating is above some threshold; otherwise, it is under a control treatment $t = 0$. For the outcome, $y = 1$ means a restaurant is completely booked and $y = 0$ means it is not.

2 PRELIMINARIES

To build a solid background to tackle the challenges of learning causality with massive data, we present the preliminaries for both structural causal models and the potential outcome framework. First, we need serious representations for causal knowledge, which are often referred to as the *causal models*. A causal model is a mathematical abstract that quantitatively describes the causal relationships between variables. *No causes in, no causes out*, the quote from Cartwright [25] summarizes the procedure of learning causality with data. First, causal assumptions or prior causal knowledge can be represented by an incomplete causal model. Then, what is missing can be learned from data. The two most well-known causal models are the structural causal models (SCMs) [104] and the potential outcome framework [99, 121]. They are considered as the foundation of causality because they enable a consistent representation of prior causal knowledge, assumptions, and estimates such that we can start from the knowns (knowledge and assumptions) to learn the unknowns[10].

We present the terminologies and notations that are used throughout this survey. Table 2 shows a nomenclature. In this survey, the a lowercase letter (e.g., x) denotes a value or the corresponding random variable (RV). Bold lowercase letters denote vectors or sets (e.g., $\boldsymbol{x}$) and bold uppercase letters signify matrices (e.g., $\boldsymbol{X}$). Calligraphic uppercase letters such can signify special sets such as sets of nodes $\mathcal{V}$ and edges $\mathcal{E}$ in a graph G. $\boldsymbol{X}$ and $\boldsymbol{x}_i$ present features for all instances and that for the i-th instance, respectively. Without specification, the subscripts denote the instance and the dimension. For example, $\boldsymbol{x}_i$ denotes features of the i-th instance and $x_{,j}$ signifies the j-th feature. t denotes the treatment variable, in this work, it is often assumed to be binary and univariate. y is referred to as the outcome variable. We use the subscript and superscript of y to signify the instance and the treatment it corresponds to. For example, when the treatment is binary, y_i^1 denotes the outcome when the instance i is treated ($t_i = 1$). τ denotes a certain type of treatment effect.

2.1 Structural Causal Models

SCMs are developed toward a comprehensive theory of causation [103]. A causal model by SCMs consists of two components: the *causal graph* (causal diagram) and the *structural equations*[11].

Causal Graphs. A causal graph forms a special class of Bayesian network with edges representing the causal effect, thus it inherits the well defined conditional independence criteria.

Definition 1. Causal Graph. *A causal graph $G = (\mathcal{V}, \mathcal{E})$ is a directed graph that describes the causal effects between variables, where $\mathcal{V}$ is the node set and $\mathcal{E}$ the edge set. In a causal graph, each*

[10] As in the problem of *learning causal relationships*, we start with no causal knowledge and learn them from data. We will discuss more details about this in Section 4.

[11] The terminology structural equation model was used to denote linear equations with causal effect as the coefficient for the treatment variable. However, researchers start to use structural equations to refer to non-linear equations with a more generalized definition of causal effect [103, 110].

Table 2. Nomenclature

Nomenclature		
Terminology	Alternatives	Explanation
causality	causal relationship, causation	causal relationship between variables
causal effect		the strength of a causal relationship
instance	unit, sample, example	an independent unit of the population
features	covariates, observables pre-treatment variables	variables describing instances
learning causal effects	forward causal inference forward causal reasoning	identification and estimation of causal effects
learning causal relationships	causal discovery causal learning causal search	inferring causal graphs from data
causal graph	causal diagram	a graph with variables as nodes and causality as edges
confounder	confounding variable	a variable causally influences both treatment and outcome

node represents a random variable including the treatment, the outcome, other observed and unobserved variables. A directed edge $x \longrightarrow y$ denotes a causal effect of x on y.

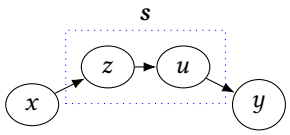
A *path* is a sequence of directed edges and a *directed path* is a path whose edges point to the same direction. In this work, we only consider *directed acyclic graphs* (DAGs) where no directed path starts and terminates at the same node. Given a SCM, the conditional independence embedded in its causal graph provides sufficient information confirm whether it satisfies the criteria such that we can apply certain causal inference methods. To understand the conditional independence, here, we briefly review the concept of *dependency-separation* (d-separation) based on the definition of *blocked* path. Fig. 1 shows three typical DAGs. In the *chain* (Fig. 1a), x causally affects y through its influence on z. In the *fork* (Fig. 1b), z is the common cause of both x and y. In this case, x is associated with y but there is no causation between them. When z is a *collider* node (see Fig. 1c), both x and y cause z but there is no causal effect or association between x and y. In the chain and fork, the path between x and y is blocked if we condition on z, which can be denoted as $x \perp\!\!\!\perp y|z$. Contrarily, in the collider (Fig. 1c), conditioning on z introduces an association between x and y, which can be represented by $x \perp\!\!\!\perp y, x \not\perp\!\!\!\perp y|z$. Generally, conditioning on a set of nodes blocks a path p iff there exists at least one node Z in p that is blocked.

Definition 2. Blocked. *We say a node z is blocked by conditioning on a set of nodes $\boldsymbol{s}$ if one of the two conditions is satisfied: (1) $z \in \boldsymbol{s}$ and z is not a collider node (Fig. 2a); (2) z is a collider node, $z \notin \boldsymbol{s}$ and no descendant of z is in $\boldsymbol{s}$ (Fig. 2b).*

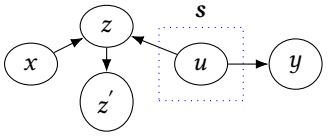
With this definition, we say a set of nodes $\boldsymbol{s}$ d-separates two variables x and y iff $\boldsymbol{s}$ blocks all paths between them. The concept of d-separation plays a crucial role in explaining causal concepts. *Causal Markovian condition* is often applied to SCMs, which means we can factorize the joint distribution represented by a *Markovian* SCM of variables $\mathcal{V} = \{x_{,1}, ..., x_{,J}\}$ with:

$$P(x_{,1}, ..., x_{,J}) = \prod_{j=1}^{J} P(x_{,j}|\boldsymbol{Pa}_{,j}, \epsilon_{,j}), \tag{1}$$

where $\boldsymbol{Pa}_{,j}$ denotes the set of parent variables of $x_{,j}$, each of which has an arrow in $x_{,j}$. Moreover, $\epsilon_{,j}$ stands for the noise term representing the causal effect of unobserved variables on $x_{,j}$. In this work,

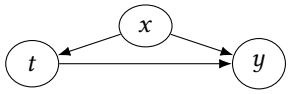


(a) Conditioning on s blocks the node z as $z \in s$ and z is not a collider.

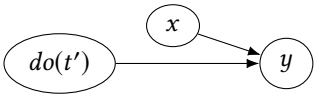


(b) Conditioning on s blocks z as z is a collider and neither z nor $z^{'}$ is in s.

Fig. 2. Examples of z being blocked by conditioning on s



(a) A SCM without intervention.



(b) A SCM under the intervention $do(t)$.

Fig. 3. SCMs without and under intervention $do(t')$ for the Yelp example, where x, t and y denote restaurant category, Yelp rating and customer flow.

we focus on Markovian SCMs. Here, we introduce the key concepts of learning causality through a toy SCM which embeds prior causal knowledge for the Yelp example [6]. In Fig. 3a, there are three random variables, namely x, t and y, which stand for the restaurant category (*confounder*), Yelp star rating (treatment) and customer flow (outcome). With the three directed edges, this causal graph embeds the knowledge of the three causal effects:

(1) The category of a restaurant influences its Yelp rating. For example, the average rating of fast food restaurants is lower than that of Ramen restaurants.
(2) The category of a restaurant also influences its customer flow. For example, the average customer flow of fast food restaurant is higher than that of Ramen restaurants.
(3) Yelp rating of a restaurant influences its customer flow, which is a common understanding.

Structural Equations. Given a causal graph, a set of equations called structural equations can be developed to specify the causal effects represented by the directed edges in the graph. We can use a set of non-parametric structural equations as the representation for three causal effects embedded in the causal graph (Fig. 3a). Specifically, this associated structural equation model can be written down as a set of equations below, where each equation corresponds to one edge in the graph:

$$x = f_x(\epsilon_x),\ t = f_t(x, \epsilon_t),\ y = f_y(x, t, \epsilon_y). \tag{2}$$

In Eq. 2, ϵ_x, ϵ_t and ϵ_y denote the noise of the observed variables. They are often assumed to be *exogenous*, and as a result, they are independent of each other. Semantically, the noise terms represent the causal effect of unobserved variables on the variable on the LHS. It is worth mentioning that for each equation, we assume that the variables on the RHS influences those on the LHS, not the other way around. Rewriting this equation in a different order as $x = f_t^{-1}(t, \epsilon_t)$ can be misleading as it embeds the causal knowledge that Yelp rating causally influences restaurant type. Thus, the causal direction is flipped, which is not desired. The structural equation (Eq. 2) provides a quantitative way to represent *intervention* on a variable of the corresponding causal graph (Fig. 3a). In SCMs, *do-calculus* [104] is developed to define intervention. Specifically, do-calculus introduces a new operator $do(t')$, which denotes setting the value of the variable t to t'. The notation of $do(t)$ leads to a formal expression of the interventional distributions as follows:

Definition 3. Interventional Distribution (Post-intervention Distribution). *The interventional distribution $P(y|do(x'))$ denotes the distribution of the variable y when we rerun the modified data-generation process where the value of variable x is set to x'.*

For example, for the causal graph in Fig. 3a, the post-intervention distribution $P(y|do(t))$ refers to the distribution of customer flow y as if the rating t is set to t' by intervention, where all the arrows into t are removed, as shown in Fig. 3b. The structural equations associated with Fig. 3b under the intervention on the treatment variable, denoted by $do(t')$, can be written as:

$$x = f_x(\epsilon_x),\ t = t',\ y = f_y(x, t, \epsilon_y), \tag{3}$$

which formulates the interventional distribution as $P(y|do(t')) = f_y(x, t, \epsilon_y)$. Then, when it comes to the causal effect of t on y, in the language of SCMs, the problem of calculating causal effects can be translated into queries about the interventional distribution $P(y|do(t))$ with different t. Implicitly, we assume that the variables follow the same causal relationships of a SCM for each instance. Hence, SCMs enable us to define *average treatment effect* (ATE). For the running example, the ATE of Yelp rating can be defined as a function:

$$\tau(t, c) = \mathbb{E}[y|do(t)] - \mathbb{E}[y|do(c)], t > c, \tag{4}$$

where t and c refer to the ratings that are considered as positive and negative, respectively. In many cases, the treatment variable is binary, thus the ATE reduces to a value $\mathbb{E}[y|do(1)] - \mathbb{E}[y|do(0)]$. However, the gap between an interventional distribution and the relevant probability (e.g., $P(y|do(t))$ and $P(y|t)$) impedes us from calculating ATE. In the literature, we call this gap as the *confounding bias*. We present the formal definition of confounding bias with do-calculus and SCMs.

Definition 4. Confounding Bias. *Given variables y, x, confounding bias exists for causal effect $x \rightarrow y$ iff the statistical association is not always the same as the corresponding interventional distribution, namely $P(y|x) \neq P(y|do(x))$.*

Confounding bias often results from the existence of *back-door path* (e.g., the path $t \leftarrow x \rightarrow y$ in Fig. 3a). Its formal definition is as follows:

Definition 5. Back-door Path. *Given a pair of treatment and outcome variables (t, y), we say a path connecting t and y is a back-door path for (t, y) iff it satisfies that (1) it is not a directed path; and (2) it is not blocked (it has no collider).*

An example of back-door path is the path $t \leftarrow x \rightarrow y$ in Fig. 3a. With the definition of back-door path, we can give the formal definition for a *confounder* or *confounding variable* as:

Definition 6. Confounder (Confounding Variable). *Given a pair of treatment and outcome variables (t, y), we say a variable $z \notin \{t, y\}$ is a confounder iff it is the central node of a fork and it is on a back-door path of (t, y).*

In particular, in the running example, the probability distribution $P(y|t)$ results from a mixture of the causal effect $P(y|do(t))$ and the statistical associations produced by the back-door path $t \leftarrow x \rightarrow y$, where x is the confounder. Note that neither $x \rightarrow t$ nor $x \rightarrow y$ is the causal effect we want to estimate. Estimating the causal effects we care about from observational data requires to eliminate confounding bias, and the procedure is referred to as *causal identification*.

Definition 7. Causal Identification. *We say a causal effect is identified iff the hypothetical distribution (e.g., interventional distribution) that defines the causal effect is formulated as a function of probability distributions.*

In other words, for causal identification, we need to block the back-door paths that reflect other irrelevant causal effects. Intuitively, a way to eliminate confounding bias is to estimate the causal effect within subpopulations where the instances are homogeneous w.r.t. confounding variables [103]. This corresponds to *adjustment* on variables that satisfy the *back-door criterion* for causal identification [104]. Now we present a formal definition of the back-door criterion.

Definition 8. Back-door Criterion. *Given a treatment-outcome pair (t, y), a set of features $\boldsymbol{x}$ satisfies the back-door criterion of (t, y) iff conditioning on $\boldsymbol{x}$ can block all back-door paths of (t, y).*

A set of variables that satisfies the back-door criterion is referred to as a *admissible set* or a *sufficient set*. For the running example, we are interested in the causal effect of Yelp star rating on the customer flow ($t \rightarrow y$) or equivalently the interventional distribution $P(y|do(t))$. So for causal identification, we aim to figure out a set of features that satisfies the back-door criterion for the treatment-outcome pair (t, y). For example, if restaurant category $x_{,j}$ is the only confounder for the causal effect of Yelp rating on customer flow, then $\boldsymbol{s} = \{x_{,j}\}$ satisfies the back-door criterion. There are two types of data w.r.t. the back-door criterion for causal inference. In the first type, we assume that the whole set or a subsets of the features $\boldsymbol{s}$ satisfies the back-door criterion such that by making adjustment on $\boldsymbol{s}$, $P(y|do(t))$ can be identified. We will introduce methods for learning causal effects with data of this type in Section 3.1. In the second type, other criteria are used to identify causal effects without the back-door criterion satisfied.

Confounding bias without back-door path. Confounding bias may exist when there is no back-door path. An example is a type of selection bias [20], when the causal graph is $t \rightarrow z \leftarrow x \rightarrow y$ and the dataset is collected only for instances with $z_i = 1$, then within this dataset, the estimated statistical association $P(y|t)$ can be non-zero although we know that there is no causal effect $t \rightarrow y$.

Beyond do-calculus. Although do-calculus plays an important role in the language of SCMs, it has some limitations, which mainly come from the assumption that the variables of all instance follow the same causal relationships. This implies that it is difficult to formulate individual-level hypothetical distributions with do-calculus in SCMs. Let us consider the running example, even if we could hack Yelp and replace the rounded star rating with the true average rating for all restaurants, we still cannot answer questions such as what would the customer flow for restaurant C be if we had increased its rating by 0.5 star without changing the ratings of others? In [105], Pearl refers to the hypothetical distributions for such cases which cannot be identified through interventions as *counterfactuals*. Naturally, do-calculus, the formal representation of hypothetical intervention, cannot help us formulate counterfactuals within the language of SCMs. Therefore, besides do-calculus, Pearl [103] introduced a new set of notations. For example, $P(y_d|y', d')$ denotes the probability of $Y = y$ if D had been observed with value d, given the fact that we observe $Y = y', D = d'$ in the data. In the running example, for a restaurant with rating d' and customer flow y', the counterfactual probability $P(y_d|y', d')$ stands for how likely the restaurant's customer flow would be y if we had observed its rating as d.

2.2 Potential Outcome Framework

The potential outcome framework [99, 121] is mainly applied to learning causal effect as it corresponds to a given treatment-outcome pair (D, Y). We start with defining the *potential outcome*:

Definition 9. Potential Outcome. *Given the treatment and outcome t, y, the potential outcome of the instance y_i^t, is the value of y would have taken if the instance i is under treatment t.*

Following this definition, the main challenge is that one potential outcome can be observed for each instance. Now we can define the *individual treatment effect* (ITE) as the difference between potential outcomes of a certain instance under two different treatments. Then we can extend ITE

to ATE on arbitrary populations. In many applied studies, researchers often assume about binary treatment ($t \in \{0, 1\}$), where $t = 1$ and $t = 0$ mean that an instance is under treatment and control, respectively. Then we can formally define that the ITE becomes as:

Definition 10. Individual Treatment Effect. *Assuming binary treatment, given an instance i and its potential outcomes y_i^t, the individual treatment effect is defined as $\tau_i = y_i^1 - y_i^0$.*

With this definition, we can extend it to ATE over the whole population being studied and other subpopulation average causal effects such as *conditional average treatment effect* (CATE). Earlier in this section, we have already defined ATE with do-calculus, here we show that ATE can also be formulated in the potential outcome framework. Formally, given the definition of ITE, we can formulate ATE as the expectation of ITE over the whole population $i = 1, ..., n$ as:

$$\tau = \mathbb{E}_i[\tau_i] = \mathbb{E}_i[y_i^1 - y_i^0] = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0), \tag{5}$$

The ATE on subpopulations can also be interesting for some studies. An example is the *conditional average treatment effect* (CATE) of instances with the same features, i.e., $\tau(\boldsymbol{x}) = \mathbb{E}_{i:\boldsymbol{x}_i=\boldsymbol{x}}[\tau_i]$.

Three assumptions of the potential outcome framework. There are three assumptions made to formulate ITE: *the stable unit treatment value assumption* (SUTVA), *consistency* and *ignorability (unconfoundedness)*. One can break down SUTVA into two conditions: *well-defined treatment levels* and *no interference*. The condition of well-defined treatment indicates that given two different instances $i \neq j$, if the values of their treatment variable are equivalent, then they receive the same treatment. The condition of no interference signifies that the potential outcomes of an instance is independent of what treatments the other units receive, which can be formally expressed as $y_i^t = y_i^{t_i}$, where $\boldsymbol{t} \in \{0, 1\}^n$ denotes the vector of treatments for all instances. The SUTVA assumption is in accordance with the implicit assumption of SCMs, where the same SCM describes the causal relationships for all instance and there is no interference between any pair of instances [105]. Although the condition of no interference is often assumed, there are cases when the inter-dependence between instances matters. For example, with *spillover effect* or *treatment entanglement*, treatment of an instance may causally influence the outcomes of its neighbors in a given graph connecting instances [11, 114, 138]. The second assumption, consistency, means that the observed outcome is independent of the how the treatment is assigned. Finally, with unconfoundedness (ignorability), we assume that all the confounding variables are observed and reliably measured by a set of features $\boldsymbol{s}$ for each instance. In SCMs, this means the set of features $\boldsymbol{s}$ satisfies the back-door criterion [103]. It is worth mentioning that here we only consider the case where confounding bias results back-door paths. *unconfoundedness* means that the values of the potential outcomes are independent of the observed treatment, given the set of confounding variables. Mathematically, unconfoundedness can be formulated as:

$$y_i^1, y_i^0 \perp\!\!\!\perp t_i | \boldsymbol{s}, \tag{6}$$

where $\boldsymbol{s}$ denotes a vector of confounders, namely a subset of features that causally influences both the treatment t_i and the outcome y_i^t. We can see that this is also an assumption defined at the individual level. Unconfoundedness directly leads to causal identification as Pearl [103] showed that ,given Eq. 6, $\boldsymbol{s}$ always satisfies the back-door criterion of (t, y). An extra condition $P(t = 1|\boldsymbol{x}) \in (0, 1)$ if $P(\boldsymbol{x}) > 0$ is usually added to make it *strong ignorability*.

At the end, we compare the two formal frameworks. The two formal frameworks are logically equivalent, which means an assumption in one can always be translated to its counterpart in the other [103]. However, there are some differences between them. In the potential outcome framework, the causal effects of the variables other than the treatment and the special variables such as

instrumental variable are not defined. This is a strength of this framework as we can model the interesting causal effects without knowing the complete causal graph [4]. While in SCMs, we are able to study the causal effect of any variable. Therefore, to learn causal relationships among an arbitrary set of variables, SCMs are often preferred [4].

3 LEARNING CAUSAL EFFECTS

In this section, we introduce methods for learning causal effects. We aim to understand how to quantify causal effects in a data-driven way. We first define the problem of learning causal effects.

Definition 11. Learning Causal Effects *Given n instances,* $[(\boldsymbol{x}_1, t_1, y_1), ..., (\boldsymbol{x}_n, t_n, y_n)]$*, learning causal effects quantifies how the outcome* y *is expected to change if we modify the treatment from* c *to* t*, which can be defined as* $\mathbb{E}[y|t] - \mathbb{E}[y|c]$*, where* t *and* c *denote a treatment and the control.*

Depending the problem, we may care about the causal effect for different populations. It can be the whole population, a known subpopulation that is defined by some conditions, an unknown subpopulation or an individual. Among all kinds of treatment effects, the average treatment effect (ATE) is often interesting when it comes to making decision on whether a treatment should be introduced to the population. Furthermore, in SCMs and do-calculus, identification of ATE only requires to query interventional distributions but not counterfactuals. This means that ATE is often easier to identify and estimate than other treatment effects. In terms of evaluation, regression error metrics such as mean absolute error (MAE) can be used to evaluate models for learning ATE. Given the ground truth τ and the inferred ATE $\hat{\tau}$, the MAE on ATE is:

$$\epsilon_{MAE\_ATE} = |\tau - \hat{\tau}|. \tag{7}$$

Similarly, we can apply MAE for average treatment effects over subpopulations.

However, when the population consists of heterogeneous groups, ATE can be misleading. For example, Yelp rating may matter much more for restaurants in big cities than those in small towns. Therefore, ATE can be spurious as an average of heterogeneous causal effects. In contrast, the average should be taken within each homogeneous group. In many cases, without knowledge about the affiliation of groups, an assumption we can make is that each subpopulation is defined by different feature values. Thus, we can learn a function to map the features that define a subpopulation to its estimated ATE. With this assumption, given a certain value of features $\boldsymbol{x}$ and binary treatment t, the CATE is a function of $\boldsymbol{x}$ and is defined as:

$$\tau(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}, t = 1] - \mathbb{E}[y|\boldsymbol{x}, t = 0]. \tag{8}$$

In this case, we assume that only the features and the treatment are two factors that determine the outcome. The target is to learn a function $\hat{\tau}$ to estimate CATE. Empirically, with cross-validation, we can evaluate the quality of the learned function $\hat{\tau}(\boldsymbol{x})$ based on the mean squared error (MSE):

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0 - \hat{\tau}(\boldsymbol{x}_i))^2, \tag{9}$$

which is often referred to as *precision in estimation of heterogeneous effect* (PEHE). It is also adopted for evaluating estimated individual treatment effects (ITE) [62, 74, 89, 126].

3.1 Learning Causal Effects with Unconfoundedness

To eliminate the confounding bias, it is often assumed that all the confounders are among the observed features. In SCMs, this is equivalent to assume that conditioning on a subset of observed features, denoted by $\boldsymbol{s}$, can block all the back-door paths for each instance. *Adjustment* eliminates confounding bias based on the subset of features $\boldsymbol{x}$. We introduce three families of adjustments:

regression adjustment, propensity score methods and covariate balancing. We assume binary treatment $t \in \{0, 1\}$ and adopt the language of generalized structural equation introduced in Section 2. The causal graph embedding the assumption for such methods is shown in Fig. 4.

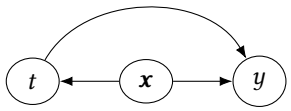


Fig. 4. A causal graph for the unconfoundedness assumption which is used for learning causal effects.

3.1.1 Regression Adjustment. In supervised learning, we fit a function to estimate the probability distribution $P(y|\boldsymbol{x})$ where y and $\boldsymbol{x}$ denote the observed label and the features. As discussed in Section 2, for learning causal effects, we are interested in interventional distributions and counterfactuals which cannot be directly estimated from data. Following the potential outcome framework, we aim to infer the counterfactual outcomes $y_i^{1-t_i}$ based on the features $\boldsymbol{x}$ and the treatment t. There are two types of regression adjustment. First, we can fit a single function to estimate $P(y|\boldsymbol{x}, t)$. This is enough for learning ITE because $\boldsymbol{x}$ is a *sufficient set*, which means by conditioning on $\boldsymbol{x}$, there would be no confounding bias, i.e. $P(y|t, \boldsymbol{x}) = P(y|do(t), \boldsymbol{x})$. So we can infer the counterfactual outcome as $\hat{y}_i^{1-t_i} = \mathbb{E}(y_i|1 - t_i, \boldsymbol{x}_i)$. In similar ways, it is also possible to fit a model for each potential outcome, i.e. $P^1(y|\boldsymbol{x}) = P(y|t = 1, \boldsymbol{x})$ and $P^0(y|\boldsymbol{x}) = P(y|t = 0, \boldsymbol{x})$. Then we can estimate ATE by:

$$\hat{\tau} = [\sum_{i=1}^{n}(\hat{y}_i^1 - \hat{y}_i^0)]/n, \tag{10}$$

where we estimate $\hat{y}_i^t$ by the model $\mathbb{E}(y|t, \boldsymbol{x}_i)$.

3.1.2 Propensity Score Methods. Propensity score methods can be considered as a special case of matching methods [97]. Matching methods divide instances into strata and treat each stratum as a randomized controlled trial (RCT). Based on this assumption, ATE is identified and can be estimated by the naïve estimator within each stratum. In matching methods, we assume *perfect stratification*, which means that (1) each group is defined by a set of features $\boldsymbol{x}$; (2) instances in a group are indistinguishable except the treatment and the potential outcomes [97]. Formally, perfect stratification means $\mathbb{E}[y_i^t|t_i = 1, f(\boldsymbol{x})] = \mathbb{E}[y_i^t|t_i = 0, f(\boldsymbol{x})], t \in \{0, 1\}$. Function $f(\boldsymbol{x})$ outputs a continuous value we stratify instances into groups based on $f(\boldsymbol{x})$. This equation can be interpreted as: given the group affiliation, the expected values of the potential outcomes do not change with the observed treatment. This is equivalent to the unconfoundedness assumption in each stratum defined by $f(\boldsymbol{x})$ whose parameterization can be flexible.

But we need to be careful when there exists a group which only contains instances with $t = 1$ or $t = 0$, where we cannot estimate ATE in such stratum with the naïve estimator. This issue is referred to as *the lack of overlap*. To deal with this problem, *matching as weighting* methods are proposed. The most widely adopted methods define the function $f(\boldsymbol{x})$ as an estimator of the *propensity score* $P(t|\boldsymbol{x})$. Although the features and treatment assignments are fixed given observational data, we assume that observed treatment is assigned by sampling from the true propensity score ($t_i \sim P(t_i|\boldsymbol{x}_i)$), a.k.a. the probability to receive treatment given the features. A natural question arises: *Does propensity score help*? Propensity score helps in the sense that it represents the features with a scalar to reduce computational cost in matching and avoid the possible data sparseness issue. Following [120], we estimate the propensity score by training a classifier to predict whether an instance would be treated given its features. $P(t|\boldsymbol{s})$ is often estimated by logistic

regression $\hat{P}(t|\boldsymbol{x};\boldsymbol{w},w_0) = \text{expit}[1+\exp(-\boldsymbol{w}^T\boldsymbol{x}-w_0)]$, where $\text{expit}(a) = \frac{1}{1+\exp(-a)}$. We can estimate the parameters by minimizing negative log-likelihood $-\frac{1}{n}\sum_{i=1}^{n}\log P(t_i|\boldsymbol{s}_i)$. Propensity score methods can be categorized into four classes [14]: *propensity score matching* (PSM), *propensity score stratification*, *inverse probability of treatment weighting* (IPTW), and *adjustment based on propensity score*. Here we focus on the PSM and IPTW as propensity score stratification is an extension of PSM, and adjustment based on propensity score is a combination of regression adjustment and propensity score methods.

Propensity Score Matching (PSM). PSM is the approach to match a treated (controlled) instance to a set of controlled (treated) instances with similar propensity scores. The most common approach is *Greedy One-to-one Matching* [54]. For each instance i, we find a matched instance j as the one with the shortest distance from i in the other treatment group. For PSM, the distance is often calculated based on the propensity scores as $dist(i,j) = |P(t|\boldsymbol{x}_i) - P(t|\boldsymbol{x}_j)|$. Once the instances are matched, we can estimate ATE as:

$$\hat{\tau} = [\sum_{i:t_i=1}(y_i - y_j) + \sum_{i:t_i=0}(y_j - y_i)]/n. \tag{11}$$

Besides the Greedy One-to-one PSM, there are many other PSM methods. The difference comes what methods we use to match instances. Readers can check [14] for various PSM methods. Stratification on propensity score is an extension of PSM. Having propensity score estimated, we can stratify instances based on the predefined thresholds on propensity scores or the number of strata.

Thus, stratum-specific ATE can be calculated by the naïve estimator. Specifically, ATE is calculated as the weighted average over all strata:

$$\hat{\tau} = \sum_j |U_j|(\frac{1}{|U_j^1|}\sum_{i\in U_j^1} y_i - \frac{1}{|U_j^0|}\sum_{i\in U_j^0} y_i)/\sum_j |U_j|, \tag{12}$$

where U_j, U_j^1 and U_j^0 denote the set of instances, treated instances and controlled instances in the i-th stratum, respectively. A combination of regression adjustment and propensity score stratification can be used to account for the difference between instances in the same stratum [14, 71, 91].

Inverse Probability of Treatment Weighting (IPTW). The IPTW [63] is a covariate balancing method. Intuitively, we can weight instances based on their propensity scores to synthesize a RCT [14]. A common way to define the sample weight w_i is by:

$$w_i = \frac{t_i}{P(t_i|\boldsymbol{x}_i)} + \frac{1-t_i}{1-P(t_i|\boldsymbol{x}_i)}. \tag{13}$$

With Eq. 13, we can find that for a treated instance i and a controlled instance j, $w_i = \frac{1}{P(t_i|\boldsymbol{x}_i)}$ and $w_j = \frac{1}{1-P(t_j|\boldsymbol{x}_j)}$. So the weight refers to the inverse probability of receiving the observed treatment (control). To synthesize a RCT, we need to balance the two treatment groups by weighting the treated instance 9 times as the instances under control, which is done by Eq. 13. Then we can calculate a weighted average of factual outcomes for the treatment and control groups:

$$\hat{\tau} = \frac{1}{n^1}\sum_{i:t_i=1} w_i y_i - \frac{1}{n^0}\sum_{i:t_i=0} w_i y_i, \tag{14}$$

where n^1, n^0 denote the number of instances under treatment and control. This is based on the idea that weighting the instances with inverse probability makes a synthetic RCT. Hence, a naïve estimator can be applied to estimate the ATE as in Eq. 14. Regression adjustment can also be applied to the weighted dataset to reduce the residual of the synthetic RCT [73]. Instances with propensity score close to 1 or 0 may suffer from an extremely large weight. In [61], Hernan proposed to stabilize weights to handle this issue in IPTW.

Doubly Robust Estimation (DRE) Funk et al. [44] proposed DRE as a combination of a regression adjustment $\mathbb{E}[y|t,\boldsymbol{x}]$ and another that estimates the propensity score $\mathbb{E}[t|\boldsymbol{x}]$. In fact, only one of the two underlying models needs to be correctly specified to make it an unbiased and consistent estimator of ATE. In particular, a DRE model estimates individual-level potential outcomes based on these two models as:

$$\hat{y}_i^1 = \frac{y_i t_i}{\hat{P}(t_i|\boldsymbol{x}_i)} - \frac{\tilde{y}_i^1(t_i - \hat{P}(t_i|\boldsymbol{x}_i))}{\hat{P}(t_i|\boldsymbol{x}_i)} \quad , \quad \hat{y}_i^0 = \frac{y_i(1-t_i)}{1-\hat{P}(t_i|\boldsymbol{x}_i)} - \frac{\tilde{y}_i^0(t_i - \hat{P}(t_i|\boldsymbol{x}_i))}{1-\hat{P}(t_i|\boldsymbol{x}_i)} \tag{15}$$

where $\tilde{y}_i^{t_i}$ denotes the estimated potential outcomes for the instance i with regression adjustment $\mathbb{E}[y|t,\boldsymbol{x}]$ and $\hat{P}(t_i|\boldsymbol{x}_i)$ is the estimated propensity score for the instance i. Taking a closer look at Eq. 15, we can find that the regression adjustment model is applied for the estimation of counterfactual outcomes as: $\hat{y}_i^{1-t_i} = \tilde{y}_i^{1-t_i}$, while more complicated, a mixture of the regression adjustment of propensity score models is developed for the factual outcomes. Then we can estimate ATE by taking the average over the estimated ITE for all the instances as in Eq. 10.

Targeted Maximum Likelihood Estimator (TMLE). Being more generalized than DRE, the *Targeted Maximum Likelihood Estimator* (TMLE) is proposed in [139]. We can estimate ATE with TMLE as: $\frac{1}{n}\sum_{i=1}^{n} Q_n^*(1,\boldsymbol{x}_i) - Q_n^*(0,\boldsymbol{x}_i)$. To obtain $Q_n^*(t,\boldsymbol{x})$, there are three steps: (1) We fit a model $Q_n^0(t,\boldsymbol{x})$ to estimate the factual outcome from the features and the treatment. (2) We fit a model $g(t=1,\boldsymbol{x})$ for the propensity score $P(t=1|\boldsymbol{x})$. (3) Given Q_n^0 and $g(t,\boldsymbol{x})$, we aim to find a better model targeted at minimizing the mean squared error (MSE) for the factual outcomes. Assuming $y \in [0,1]$, this can be done by learning a new outcome estimator $\bar{Q}_n^*(t,\boldsymbol{x})$ with parameters $\hat{\epsilon}_0$ and $\hat{\epsilon}_1$:

$$\bar{Q}_n^*(t,\boldsymbol{x}) = \text{expit}\left[Q_n^0(t,\boldsymbol{x})/(1-Q_n^0(t,\boldsymbol{x})) + \hat{\epsilon}_0 H_0(t,\boldsymbol{x}) + \hat{\epsilon}_1 H_1(t,\boldsymbol{x})\right], \tag{16}$$

where $\text{expit}(a) = \frac{1}{1+\exp(-a)}$, $H_0(t,\boldsymbol{x}) = -\frac{\mathbb{1}(t=0)}{g(t=0|\boldsymbol{x})}$ and, $H_1(t,\boldsymbol{x}) = \frac{\mathbb{1}(t=1)}{g(1|W)}$.

3.1.3 Covariate Balancing. Besides reweighting samples with propensity scores, the *confounding balancing* methods learn sample weights through regression [79].

Entropy Balancing (EB). Hainmueller [58] proposed EB, a preprocessing method for covariate balancing. The goal is to learn sample weights of the instances under control such that the moments of the two groups are matched. The weights are learned by minimizing the objective:

$$\arg\min_{w_i} H(\boldsymbol{w}) = \sum_{i:t_i=0} d(w_i) \; s.t. \; \sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i) = m_r \; with \; r = 1,...,R, \tag{17}$$

where $sum_{i:t_i=0} w_i = 1;\; w_i \geq 0, \forall i \in \{i|t_i = 0\}$. $d(\cdot)$ is a distance metric (e.g., KL divergence $d(w_i) = w_i \log(w_i/q_i)$) measuring the distance between the learned weights $\boldsymbol{w}$ and base weights $\boldsymbol{q}, q_i \geq 0$, and $\sum_i q_i = 1$. We can use uniform weights $q_i = 1/n^0$, where n^0 denotes the number of instances under control. $\sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i) = m_r$ refers to a set of R balance constraints where $c_{ri}(\boldsymbol{x}_i)$ is specified as a moment function for the control group and m_r denotes the counterpart of the treatment group. For example, when $c_{ri}(\boldsymbol{x}_i) = (x_i^j)^r$, then $\sum_{i:t_i=0} w_i c_{ri}(\boldsymbol{x}_i)$ denotes the reweighted r-th moment of the feature x^j for the control, and therefore, m_r would contain the r-th order moment of a feature x^j from the treatment group. Compared to other balancing methods, EB allows large set of constraints such as moments of feature distributions and interactions. In addition, different from the matching methods, EB keeps weights close to the base weights to prevent information loss.

Approximate Residual Balancing (ARB). ARB [13] combines balancing weights with a regularized regression adjustment for learning ATE from high-dimensional data. ARB consists of three

steps. First, the sample weights $\boldsymbol{w}$ are learned as:

$$\arg\min_{\boldsymbol{w}} (1-\xi)||\boldsymbol{w}||_2^2 + \xi||\frac{1}{n^1}\sum_{i:t_i=1}\boldsymbol{x}_i - \boldsymbol{X}_{i:t_i=0}^T\boldsymbol{w}||_\infty^2 \; s.t. \sum_{i:t_i=0} w_i = 1 \; and \; w_i \in [0,(n^0)^{-2/3}], \quad (18)$$

where $\boldsymbol{X}_{i:t_i=0}$ denotes the feature matrix for the control group. Then a regularized linear regression adjustment model with parameters $\boldsymbol{\beta}$ fitted as:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i:t_i=0}(y_i - \boldsymbol{x}^T\boldsymbol{\beta})^2 + \lambda((1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1), \quad (19)$$

where λ and α are hyperparameters controlling the strength of regularization. At the end, we can estimate ATE as $\hat{\tau} = \frac{1}{n^1}\sum_{i:t_i=1} y_i - (\frac{1}{n^1}\sum_{i:t_i=1}\boldsymbol{x}_i^T\boldsymbol{\beta} + \sum_{i:t_i=0} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}))$. Compared to EB [58], ARB handles sparseness of high-dimensional data with lasso and elastic net [136].

Covariate Balancing Propensity Score (CBPS). CBPS [69], a method robust to misspecification of propensity score model, is proposed to model propensity scores and balance covariate simultaneously. Assuming the propensity score model $f(\boldsymbol{x})$ with parameters $\boldsymbol{\beta}$, the efficient GMM estimator is used to learn $\boldsymbol{\beta}$:

$$\arg\min_{\boldsymbol{\beta}} [\frac{1}{n}\sum_i g(t_i,\boldsymbol{x}_i)]^T\Sigma(\boldsymbol{t},\boldsymbol{X})^{-1}[\frac{1}{n}\sum_i g(t_i,\boldsymbol{x}_i)], \quad (20)$$

where $g(t_i,\boldsymbol{x}_i) = (\frac{t_i f'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} - \frac{(1-t_i)f'(\boldsymbol{x}_i)}{1-f(\boldsymbol{x}_i)}, \frac{(t_i - f(\boldsymbol{x}_i))f'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)(1-f(\boldsymbol{x}_i))})^T$ signifies the moment conditions. They are derived from the KKT condition of minimizing $-\sum_{i=1}^n t_i \log f(\boldsymbol{x}_i) + (1-t_i)\log(1-f(\boldsymbol{x}_i))$, the MLE estimator by which we learn $\boldsymbol{\beta}$. Similar to EB [58], CBPS combines two methods: covariate balancing and IPTW. Compared to EB, CBPS directly models propensity scores.

3.2 Learning Causal Effects with Unobserved Confounders

In many real-world problems of learning causal effects, there exist unobserved confounders. In these cases, the assumption of unconfoundedness is not satisfied. In the language of SCMs, this means we are not able to block back-door path by conditioning on the features. Therefore, a family of methods are developed to handle this situation. The intuition is to utilize alternative information. Here, we focus on three most popular methods for learning causal effects with unobserved confounders: *instrumental variable methods*, *front-door criterion*, and *regression discontinuity design*.

3.2.1 Instrumental Variable Methods. *Instrumental variables* enable us to learn causal effects with unobserved confounders, which are defined as:

Definition 12. Instrumental Variable *Given an observed variable i, features $\boldsymbol{x}$, the treatment t and the outcome y, we say i is a valid instrumental variable (IV) for the causal effect of $t \to y$ iff i satisfies: (1) $i \not\perp\!\!\!\perp t|\boldsymbol{x}$, and (2) $i \perp\!\!\!\perp y|\boldsymbol{x}, do(t)$ [8].*

This means a valid IV causally influences the outcome only through affecting the treatment. In SCMs, the first condition means there is an edge $i \to t$ or a non-empty set of collider(s) $\boldsymbol{x}$ s.t. $i \to \boldsymbol{x} \leftarrow t$ where $\boldsymbol{x}$ denotes the features or a subset of features. The second condition requires that $i \to t \to y$ is the only path that starts from i and ends at y. Thus, blocking t makes $i \perp\!\!\!\perp y$. This implies the *exclusive restriction* that there must not exist direct edge $i \to y$ or path $i \to \boldsymbol{x}' \to y$ where $\boldsymbol{x}' \subseteq \boldsymbol{x}$. Mathematically, for all t and $i \neq j$, this can also be denoted by $y(do(i),t) = y(do(j),t)$.

In the running example, if we only observe one confounder - the restaurant type (x), while the other confounder ($\boldsymbol{z}$) remain unobserved. By assuming that whether a customer submits a review (i) is an exogenous random variable, then it is a valid IV (Fig. 5). This is because i causally influences t and it can only causally affect y through its influence on t. With a valid IV, we identify the causal effect $t \to y$ if both the interventional distributions - $P(t|do(i))$ and $P(y|do(i))$ are identifiable.

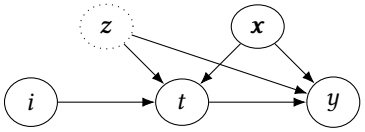


Fig. 5. A causal graph of a valid instrumental variable (i) when there are unobserved confounders (z). The binary exogenous variable i stands for whether a customer submits a review. The restaurant type (x) is an observed confounder and z is a set of unobserved confounders.

A Linear SCM for IV Estimator. Here, we show an example of IV estimator with SCMs. We assume that the structural model is linear. If we also assume that the observed and unobserved confounders $\boldsymbol{x}$ and $\boldsymbol{u}$ come with zero mean, we can write down the structural equations for the causal graph in Fig. 5 as:

$$\begin{aligned} t &= \alpha_i i + \boldsymbol{\alpha}_z^T \boldsymbol{z} + \boldsymbol{\alpha}_x^T \boldsymbol{x} + \alpha_0 + \epsilon_t \\ y &= \tau t + \boldsymbol{\beta}_z^T \boldsymbol{z} + \boldsymbol{\beta}_x^T \boldsymbol{x} + \beta_0 + \epsilon_y, \end{aligned} \tag{21}$$

where ϵ_t and ϵ_y are Gaussian noise terms with zero mean. By substituting t in the second equation with the RHS of the first equation in Eq. 21, we get:

$$y = \tau\alpha_i i + (\tau\alpha_z + \beta_z)^T \boldsymbol{z} + (\tau\alpha_x + \beta_x)^T \boldsymbol{x} + \gamma_0 + \eta, \tag{22}$$

where $\gamma_0 = \tau\alpha_0 + \beta_0$ and $\eta = \tau\epsilon_d + \epsilon_y$. Then it is not difficult to figure out an estimator for the average treatment effect (τ):

$$\hat{\tau} = (\mathbb{E}[y|i] - \mathbb{E}[y|i']) / (\mathbb{E}[t|i] - \mathbb{E}[t|i']). \tag{23}$$

Here, we rely on the following assumptions: linear structural equations, valid IV, zero-mean additive noise, and unobserved confounders. What if some of them are not satisfied in an interesting dataset? For example, the causal relationship is non-linear. In the following example, with the potential outcome framework we show this estimator also works.

An IV Estimator under the potential outcome framework. The potential outcome framework formulates the individual causal effect of the IV i on the outcome y as:

$$y_j(i_k = 1, t_l(i_k = 1)) - y_j(i_k = 0, t_l(i_j = 0)), \tag{24}$$

where $y_j(i_k, t_l)$ and $t_l(i_k)$ are the value of y and t by setting the value of the k-th IV to i_k. We also assume the IVs are binary. With the *exclusion restriction*, we know that i affects y through its influence on t, so we remove i_j that explicitly influences the value of y_j and reduces Eq. 24 to:

$$\begin{aligned} &[y_j^1 P(t_j = 1|i_j = 1) + y_j^0 P(t_j = 0|i_j = 1)] - [y_j^1 P(t_j = 1|i_j = 0) + y_j^0 P(t_j = 0|i_j = 0)] \\ &= (y_j^1 - y_j^0)(P(t_j = 1|i_j = 1) - P(t_j = 1|i_j = 0)). \end{aligned} \tag{25}$$

With the expectation over the population we obtain the same estimator as in Eq. 23. This implies that this estimator works even when causal relations are non-linear. The difficulty mainly lies in computing the influence of i on y, which is represented by the interventional distribution $P(y|do(i))$:

$$\mathbb{E}[y|do(i)] = \int_{\mathcal{T}} \mathbb{E}[y|do(t)] P(t|do(i)) dt. \tag{26}$$

You can refer to [87] for the heuristics that approximate the integral on the RHS of Eq. 26.

Two-stage Least Square (2SLS). As the IV estimator in Eq. 23 is restrictive, we may have to control a set of features $\boldsymbol{x}$ to block the back-door paths between the IV and the outcome so that the IV can be valid. These cases make it difficult or infeasible to use the estimator in Eq. 23. So we introduce 2SLS [7]. Fig. 6 shows an example for such cases where $\boldsymbol{x}$ denotes the set of confounders

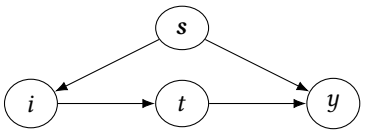


Fig. 6. Assuming that we can all confounders $\boldsymbol{x}$ for the causal effect of the IV, whether a customer writes a review on customer flow y, 2SLS can estimate the treatment effect of rating on customer flow ($t \rightarrow y$).

(e.g., whether a coupon can be found on Yelp for the restaurant) for the causal effect of whether a customer makes a review on the customer flow $i \rightarrow y$. To make i valid, the back-door path $i \leftarrow \boldsymbol{x} \rightarrow y$ has to be blocked. Besides, we may have multiple IVs for each treatment and multiple treatments. Assuming there is a set of treatments $\boldsymbol{t}$ and each treatment $t_{,j}$ has a set of IVs $\boldsymbol{i}_{,j}$. In 2SLS, two regressions are performed to learn the causal effects $(t_{,1} \rightarrow y), ..., (t_{,j} \rightarrow y), ...$: (1) we fit a function $\hat{t}_{,j} = f_{t_{,j}}(\boldsymbol{i}_{,j}, \boldsymbol{x}_{,j})$ for each treatment variable $t_{,j}$. (2) we learn a function $y = g(\hat{\boldsymbol{t}}, \boldsymbol{x})$ where $\hat{\boldsymbol{t}}$ signifies the set of treatments. Then the coefficient on $D_{,j}$ is a consistent estimate of the ATE of the j-th treatment $t_{,j}$ on y. The intuition of 2SLS follows how we find a valid IV. In the first stage, we estimate how much a certain treatment $t_{,j}$ changes if we modify the relevant IV $i_{,j}$. In the second stage, we see how the changes in $t_{,j}$ caused by $i_{,j}$ would influence y.

3.2.2 Front-door Criterion. The front-door criterion [102] enables us to learn a causal effect $t \rightarrow y$ with unobserved confounders. With front-door criterion we condition on a set of variables $\boldsymbol{m}$ which satisfies the following three conditions: (1) $\boldsymbol{m}$ blocks all the directed paths from t to y. (2) There are no unblocked back-door paths from t to $\boldsymbol{m}$. (3) t blocks all the back-door paths from $\boldsymbol{m}$ to y. In other words, we say that the set of variables $\boldsymbol{m}$ *mediates* the causal effect of t on y. From the first condition, we decompose $t \rightarrow y$ to a product of $t \rightarrow \boldsymbol{m}$ and $\boldsymbol{m} \rightarrow y$ as: $P(y|do(d)) = \int_{\mathcal{M}} P(y|do(\boldsymbol{m}))P(\boldsymbol{m}|do(d))d\boldsymbol{m}$. The second condition means there is no confounding bias for the causal effect of t on $\boldsymbol{m}$:

$$P(\boldsymbol{m}|do(d)) = P(\boldsymbol{m}|d). \tag{27}$$

The third condition allows us to infer $P(y|do(\boldsymbol{m}))$ by:

$$P(y|do(\boldsymbol{m})) = \int_{\mathcal{T}} P(y|t, \boldsymbol{m})P(t)dt. \tag{28}$$

Then the interventional distribution corresponding to $t \rightarrow y$ can be identified as:

$$P(y|do(d)) = \int_{\mathcal{M}} P(\boldsymbol{m}|d) \sum_{t \in \mathcal{T}} P(y|t, \boldsymbol{m})P(t). \tag{29}$$

Machine learning models can be applied to estimating the probabilities on the RHS of Eq. 29 from observational data. For example, assuming that the set of variables $\boldsymbol{m}$ represents the ranking of a restaurant in the search results. When the ranking is decided by the Yelp rating, $(\boldsymbol{z} \perp\!\!\!\perp \boldsymbol{x}|t, y)$, $\boldsymbol{m}$ satisfies the front-door criterion (Fig. 7a). However, when the ranking $\boldsymbol{m}$ is affected by both the rating t and confounders $\boldsymbol{z}$ (e.g. the restaurant category), then $\boldsymbol{m}$ is not a valid set of mediators (Fig. 7b). Different from the back-door criterion, the front-door criterion enables us to learn causal effects when some confounders are unobserved.

3.2.3 Regression Discontinuity Design. Sometimes, treatment assignments may only depend on the value of a special feature, which is the *running variable* r. For example, the treatment is determined by whether its running variable is greater than a cut-off value r_0. The study of the causal effect of Yelp star rating r on the customer flow y is a perfect example for such a case [6]. Yelp shows the rating of a restaurant rounded to the nearest half star. For example, restaurant i with

(a) $\boldsymbol{m}$ satisfies the front-door criterion. (b) $\boldsymbol{m}$ violates the front-door criterion.

Fig. 7. Two causal graphs where $\boldsymbol{m}$ satisfies and violates the front-door criterion

average rating 3.26 and restaurant j with 3.24 would be shown with 3.5 and 3.0 stars. Based on this fact, we can say $r_0 = 3.25$ is a cut-off which defines the treatment variable. Then for a restaurant with average rating $R \in [3, 3.5]$, we say it receives treatment ($D = 1$) when its rounded star rating is greater than its average rating ($R \geq r_0$). Otherwise, we say a restaurant is under control ($D = 0$). The intuition for Sharp Regression Discontinuity Design (Sharp RDD) [6, 23] is that the restaurants with average rating close to the cutoff $r_0 = 3.25$ are homogeneous w.r.t. the confounders. Therefore, what can make a difference in their factual outcomes is the treatment. In other words, the treatments are randomly assigned to such restaurants, which lead to the identification of the ATE. In Sharp RDD, we assume that the observed outcome is a function of the running variable as:

$$y_i = f(r_i) + \tau t_i + \epsilon_i = f(r_i) + \tau \mathbb{1}(r_i \geq r_0) + \epsilon_{yi}, \tag{30}$$

where $f(\cdot)$ is a function which is continuous at r_0, τ is the ATE and ϵ_{yu} denotes the noise term. The choice of function $f(\cdot)$ can be flexible. But the risk of misspecification of $f(\cdot)$ exists. For example, Gelman and Imbens [47] pointed out that high-order polynomials can be misleading in RDD. In the Yelp study, the fact that customers' decision on which restaurant to go solely relies on the Yelp rating supports this assumption. For many other real-world problems, however, it is not always the case where we can obtain a perfect cutoff value like the Yelp rating $r_0 = 3.25$ (stars) and the minimum drinking age $r_0 = 21$ (years old) [24]. The *Fuzzy RDD* method [9, 23] is developed to handle the cases when cut-offs on the running variable are not strictly implemented. For example, users may see the real average rating when they look into details of the restaurants and find out that the two restaurants i and j are not that different in terms of rating. Similar to the propensity score methods, Fuzzy RDD assumes the existence of a stochastic treatment assignment process $P(t|r)$. But $P(D|R)$ is also assumed to be discontinuous. The structural equations for Fuzzy RDD is:

$$\begin{aligned} y_i &= f(r_i) + \tau t_i + \epsilon_{yi} = f(r_i) + \pi_2 \mathbb{1}(r_i > r_0) + \epsilon'_{yi} \\ t_i &= g(r_i) + \pi_1 \mathbb{1}(r_i > r_0) + \epsilon_{ti} \end{aligned} \tag{31}$$

where $\tau = \frac{\pi_2}{\pi_1}$ is the ATE we want to estimate, ϵ_{yi}, ϵ'_{yi} and ϵ_{ti} are the noise terms. As τ is a division between the causal effects $\mathbb{1}(r > r_0) \rightarrow t$ and $\mathbb{1}(r > r_0) \rightarrow y$, Fuzzy RDD can be considered as an IV method where the discontinuous variable $\mathbb{1}(r > r_0)$ plays the role of IV.

4 CAUSAL DISCOVERY: LEARNING CAUSAL RELATIONSHIPS

In learning causal relationship (a.k.a. causal discovery), it examines whether a causal relationship exists. We define this problem as:

Definition 13. Learning Causal Relationships. *Given J variables, $\{\boldsymbol{x}_{,j}\}_{j=1}^{J}$, we aim to determine whether the j-th variable $\boldsymbol{x}_{,j}$ changes if we modify the j'-th variable $\boldsymbol{x}_{,j'}$ for all $j \neq j'$.*

In the running example, learning causal relationships enable us to answer the questions such as: *Do other features such as location also causally affect the customer flow? Is location a confounder for the causal effect of Yelp rating on customer flow?*To achieve this, we postulate that causality can be detected amongst statistical dependencies [110, 122]. An algorithm solving this problem

(a) A graph has $x_{,2} \perp\!\!\!\perp x_{,3}|x_{,1}$ (b) A graph has $x_{,2} \perp\!\!\!\perp x_{,3}|x_{,1}$

Fig. 8. Two exemplary causal graphs that belong to an equivalence class

learns a set of causal graphs as candidates [131]. To evaluate the learned causal relationships, we often compare each of the learned causal graphs $\hat{G}$ with the ground-truth G. The concept of the *equivalence class* is important for comparing different causal graphs.

Definition 14. Equivalence Class. *We say that two causal graphs G and G' belong to the same equivalence class iff each conditional independence that G has is also implied by G' and vise versa.*

Fig. 8a and 8b show two causal graphs which belong to the same equivalence class. Both of them have the same set of conditional independence $\{x_{,2} \perp\!\!\!\perp x_{,3}|x_{,1}\}$. A simple metric is the number of learned graphs that are equivalent to the ground truth G. In some work, we can also count the number of learned causal graphs G' that have G as a subgraph [28]. In [130], the distance between the adjacency matrix of a learned causal graph ($\hat{\boldsymbol{A}}$) and that of the ground truth ($\boldsymbol{A}$) is measured by the Frobenius norm.

4.1 Learning Causal Relationships with i.i.d. Data

Constraint-based Algorithms. This class of algorithms learn a set of causal graphs which satisfy the conditional independence embedded in the data. These algorithms use statistical tests to verify if a candidate graph fits all the independence based on the *faithfulness* assumption [131]:

Definition 15. Faithfulness. *Conditional independence between a pair of variables, $x_{,j} \perp\!\!\!\perp x_{,j'}|\boldsymbol{z}$ for $x_{,j} \neq x_{,j'}, \boldsymbol{z} \subseteq \boldsymbol{x} \setminus \{x_{,j}, x_{,j'}\}$, can be estimated from a dataset $\boldsymbol{X}$ iff $\boldsymbol{z}$ d-separates $x_{,j}$ and $x_{,j'}$ in the causal graph $G = (\mathcal{V}, \mathcal{E})$ which defines the data-generation process for $\boldsymbol{X}$.*

Under this assumption, in Fig. 4, rating is a dependent of the customer flow ($t \not\perp\!\!\!\perp y$). The challenge is mainly the computational cost as the number of possible causal graphs is super-exponential to the number of variables. Hence, algorithms are proposed to reduce the number of tests.

The Peter-Clark (PC) Algorithm. The PC algorithm [131] works in a two-step fashion. First, it learns an undirected (*skeleton graph*) from data. Then, it detects the directions of the edges to return an equivalent class of causal graphs. It starts with a fully connected graph and $q = 0$. Then for each ordered pair of connected variables $(x_{,j}, x_{,j'})$, it tests if the conditional independence $x_{,j} \perp\!\!\!\perp x_{,j'}|\tilde{\boldsymbol{z}}$ is satisfied for each $\tilde{\boldsymbol{z}} \subseteq \mathcal{N}(x_{,j})$ or $\tilde{\boldsymbol{z}} \subseteq \mathcal{N}(x_{,j'})$ of size q, where $\mathcal{N}(\cdot)$ denotes the set of neighbors of a variable. If the conditional independence is satisfied, it removes the edge $(x_{,j}, x_{,j'})$ and saves $\tilde{\boldsymbol{z}}$ as the separating set of $(x_{,j}, x_{,j'})$. Once all such edges are removed, the depth increases by 1 and this process continues till the number of neighbors for each variable is less than q. In the second step, we decide the directions of edges. We first determine *v-structures*. For a triple $(x_{,j}, x_{,j'}, x_{,j''})$ with no edge between $x_{,j}$ and $x_{,j''}$, we make it a v-structure $x_{,j} \rightarrow x_{,j'} \leftarrow x_{,j''}$ iff $x_{,j'} \notin \tilde{\boldsymbol{z}}$, where $\tilde{\boldsymbol{z}}$ denotes saved separating set of $x_{,j}$ and $x_{,j''}$. Then the remaining undirected edges are oriented following the three rules: (1) We orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there exists an edge $x_{,j''} \rightarrow x_{,j'}$ and $x_{,j''}$ and $x_{,j}$ are not neighbors. (2) We orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there is a chain $x_{,j} \rightarrow x_{,j''} \rightarrow x_{,j'}$. (3) We orient $x_{,j} - x_{,j'}$ to $x_{,j} \rightarrow x_{,j'}$ if there are two chains $x_{,j} - x_{,k} \rightarrow x_{,j'}$ and $x_{,j} - x_{,l} \rightarrow x_{,j'}$.

Other constraint-based algorithms include the IC algorithm [104] and their variants [75, 82]. However, most standard statistical tests require Gaussian or multinomial distributions. To overcome these restrictions, novel conditional independence tests are proposed for more flexible distributions [43, 116, 124, 147]. To take unobserved confounders into consideration, algorithms

are introduced to search through an extended space of causal graphs such as FCI and its extensions [32, 132]. Moreover, to go beyond observational data, Kocaoglu et al. [77] considered the problem of designing a set of interventions with minimum cost to uniquely identify any causal graph from the given skeleton. They showed the problem can be solved in polynomial time.

There are two main drawbacks of this family of algorithms. First, the faithfulness assumption can be violated by data with limited samples where independence tests may even contradict each other. Second, it may not tell causal direction between two variables.

Score-based Algorithms. To relax the faithfulness assumption, score-based algorithms replace conditional independence tests with the goodness of fit tests. Score-based algorithms learn causal graphs by maximizing the scoring criterion $S(\boldsymbol{X}, G')$ which returns the score of the causal graph G' given data $\boldsymbol{X}$. Intuitively, low scores should be assigned to the graphs which embed incorrect conditional independence. For goodness of fit tests, two components need to be specified: the structural equations and the score function. First, we consider the structural equations. Structural equations are often assumed to be linear with additive Gaussian noise [28], which introduces parameters $\boldsymbol{\theta}$. Each structural equation describes how a variable is causally influenced by its parent variables and a noise term. The second component is a score function which maps a candidate causal graph to a scalar based given a certain parameterization of structural equations. The Bayesian Information Criterion (BIC) score [123] is the most widely adopted metric $S(\boldsymbol{X}, G') = \log P(\boldsymbol{X}|\hat{\boldsymbol{\theta}}, G') - \frac{J}{2}\log(n)$, where $\hat{\boldsymbol{\theta}}$ is the MLE of the parameters, J denotes the number of variables and n signifies the number of instances. BIC score prefers causal graphs that can maximize the likelihood of observing the data with regularization on the number of parameters and the sample size. In [119], a similar score function is proposed based on maximum likelihood estimation with a different regularizer. Moreover, from the Bayesian perspective, with priors over causal graph structure and parameters, posteriors can be used to define scores. For example, Bayesian Dirichlet score [59] assumes Dirichlet prior on parameters for the multinomial distributions of variables. With the two components fixed, score of a certain causal graph for a given dataset is well defined. Then we focus on searching for the causal graphs which provide the best score for a given dataset. Searching for the causal graph with maximal score, also known as structural learning is both NP-hard and NP-complete [27, 29]. It is not computationally feasible to score all possible causal graphs exhaustively. Therefore, heuristics such as GES [28] and its extension, Fast GES (FGES) [115] are proposed to reach a locally optimal solution. When it comes to interventional data, Wang et al. [142] proposed algorithms to learn causal relationships when a mixture of interventional and observational data is given, which are non-parametric and handle non-Gaussian data well.

Greedy Equivalence Search (GES). Here we introduce GES as an example of score-based algorithms. In [28], assuming discrete variables, the BDeu criterion is used:

$$S_{BDeu}(G', \boldsymbol{X}) = \log \prod_{j=1}^{J} 0.001^{(r_j-1)q_j} \prod_{k=1}^{q_j} \frac{\Gamma(10/q_j)}{\Gamma(10/q_j + N_{jk})} \prod_{l=1}^{r_j} \frac{\Gamma(10/(r_i q_i) + N_{jkl})}{\Gamma(10/(r_j q_j))}, \tag{32}$$

where r_j and q_j signify the number of configurations of variable $x_{\cdot j}$ and parent set $\boldsymbol{Pa}_j$ specified by the graph G'. $\Gamma(n) = (n-1)!$ is the Gamma function. N_{jkl} denotes the number of records for which $x_{\cdot j} = k$ and $\boldsymbol{Pa}_j$ is in the k-th configuration and $N_{jk} = \sum_l N_{jkl}$. After initialized with the equivalent class of DAG models with no edges, two stages of greedy search are performed. First, a greedy search is performed only to insert edges. The insertion operator $Insert(x_{\cdot j}, x_{\cdot j'}, \boldsymbol{z})$ takes three inputs. $x_{\cdot j}$ and $x_{\cdot j'}$ are non-adjacent nodes in the current graph, $\boldsymbol{z}$ denotes any subset of $x_{\cdot j'}$'s neighbors that are not adjacent to $x_{\cdot j}$. The insertion operator modifies the graph by (1) adding the edge $x_{\cdot j} \rightarrow x_{\cdot j'}$ and (2) directing the previous undirected edge $z - x_{\cdot j'}$ as $z \rightarrow x_{\cdot j'}$. It is worth mentioning that undirected edges can result from the equivalent class of graphs. As a

greedy algorithm, in each iteration, for the current graph, we find the triple $x_{,j}, x_{,j'}, \boldsymbol{z}$ leading to the best score (Eq. 32) and perform the insert operator until a local maximum is reached. Then, the second greedy search is performed initialized with the local optimum of the previous phase, only to delete edges. The delete operator, $Delete(x_{,j}, x_{,j'}, \boldsymbol{z})$ takes two adjacent nodes $x_{,j}$ and $x_{,j'}$ with edge $x_{,j} - x_{,j'}$ or $x_{,j} \rightarrow x_{,j'}$ and $\boldsymbol{z}$ denoting any subset of neighbors of $x_{,j'}$ which are also adjacent to $x_{,j}$. Again, for each iteration, given the current graph, the triple $x_{,j}, x_{,j'}, \boldsymbol{z}$ with the highest score would be selected to update the graph using the delete operator. Finally, GES terminates with the local maximum reached by the second phase.

Algorithms based on Functional Causal Models (FCMs). In FCMs, a variable $x_{,j}$ can be written as a function of its directed causes $\boldsymbol{Pa}_j$ and some noise term ϵ_j as $x_{,j} = f(\boldsymbol{Pa}_j, \epsilon_j)$. Different from the two families of methods mentioned above, with FCMs, we are able to distinguish between different DAGs from the same equivalent class. Here, we adopt Linear Non-Gaussian Acyclic Model (LiNGAM) [129] as the FCM to introduce algorithms with FCMs. In the matrix form, the LiNGAM model can be written as:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{33}$$

where $\boldsymbol{x}$, $\boldsymbol{A}$ and $\boldsymbol{\epsilon}$ denote the vector of variables, the adjacency matrix of the causal graph [128], and the vector of noise, respectively. Columns of both $\boldsymbol{x}$ and $\boldsymbol{A}$ are sorted according to the *causal order* ($k(j)$) of each variable, respectively. In the LiNGAM model, the task of learning causal relationships turns into estimating a strictly lower triangle matrix $\boldsymbol{A}$ which determines a unique causal order $k(j)$ for each variable $x_{,j}$. For example, if a FCM can be specified by a LiNGAM as follows:

$$\begin{bmatrix} s \\ d \\ y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1.2 & 0 & 0 \\ 0.8 & 1.3 & 0 \end{bmatrix} \begin{bmatrix} s \\ d \\ y \end{bmatrix} + \begin{bmatrix} \epsilon_s \\ \epsilon_d \\ \epsilon_y \end{bmatrix}, \tag{34}$$

then the causal order of the three variables s, d, y is 1, 2 and 3, respectively.

ICA-LiNGAM. Based on independent component analysis (ICA) [67], the ICA-LiNGAM algorithm [129] is proposed to learn causal relationships with the LiNGAM model, with which we estimate the matrix $\boldsymbol{A}$. First, we can rewrite Eq. 33 as $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{\epsilon}$, where $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{A})^{-1}$. As each dimension of $\boldsymbol{\epsilon}$ is assumed to be independent and non-Gaussian, it defines the ICA model for the LiNGAM. Thus we can apply ICA to obtain an estimate of $\boldsymbol{B}$. Specifically, given data $\boldsymbol{X}$ of the variables $\boldsymbol{x}$, we use ICA algorithm [67] to obtain the decomposition $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{S}$. We can learn $\boldsymbol{W} = \boldsymbol{B}^{-1}$ by maximizing the objective:

$$\sum_{j=1}^{J} J_G(\boldsymbol{w}_j) \; s.t. \; \mathbb{E}[(\boldsymbol{w}_k^T \boldsymbol{x})(\boldsymbol{w}_l^T \boldsymbol{x})] = \delta_{kl}, \tag{35}$$

where $J_G(\boldsymbol{w}_i) = \{\mathbb{E}[G(\boldsymbol{w}_i^T \boldsymbol{x})] - \mathbb{E}[G(\boldsymbol{v})]\}^2$, G can be any nonquadratic function (e.g., $G(y) = y^4$). $\boldsymbol{v}$ denotes samples from a normal distribution $\mathcal{N}(0, 1)$ and δ_{kl} is the magnitude of dependence between the two variables. Then an initial estimate of $\boldsymbol{A}$, namely $\boldsymbol{A}'$, is computed based on $\boldsymbol{W}$ as $\boldsymbol{A}' = \boldsymbol{I} - \tilde{\boldsymbol{W}}'$. $\tilde{\boldsymbol{W}}'$ is obtained by dividing each row of $\tilde{\boldsymbol{W}}$ by the corresponding diagonal element. $\tilde{\boldsymbol{W}}$ is calculated by finding the unique permutation of rows of $\boldsymbol{W}$ which is nonzero on the diagonal. Finally, to estimate the causal order $k(j)$ for each $x_{,j}$, permutations are applied to $\boldsymbol{A}'$ to obtain an estimate of $\boldsymbol{A}$ which is as close to a strictly lower triangle matrix as possible. A main downfall of ICA-LiNGAM is that ICA algorithms may converge to local optima. To guarantee the convergence to the global optima in a fixed number of steps, Shimizu et al. proposed the DirectLiNGAM algorithm [130], which also determines $\boldsymbol{A}$ through estimating the causal ordering of variables $k(j)$.

Recently, Additive Noise Models (ANMs) are proposed to relax the linear restriction on the relationships between variables and the distribution of noise [64, 65]. ANMs also help reduce the

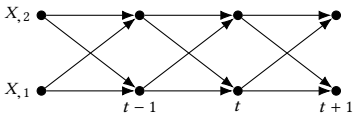


Fig. 9. An example of a chain causal graph for time series

search space of causal graph as data normally does not admit two ANMs with conflicts in directions of causal effects [65, 110]. One step further, Post-nonlinear Models expand the functional space with non-linear relationships between the variables and the noise [146].

4.2 Learning Causal Relationships with Time Series Data

Time series is an important type of non-i.i.d. data for influential applications such as speech recognition [53] and quantitative trading [5]. As causality can be implied by the arrow of time [112], it can have various definitions for time series data. Practically, researchers studied *Granger causality*[12] [52] to approximate real causality as it does not require a pre-defined causal model [16, 17, 39, 90]. In the matrix form, the linear Granger causality can be written as:

$$\boldsymbol{X}(l) = \boldsymbol{A}\boldsymbol{X}(l-1) + \boldsymbol{\epsilon}(l-1), \tag{36}$$

where the matrix $\boldsymbol{A}$ contains the temporal causal relations. Although Granger causality is merely temporal constrained statistical association, under the faithfulness assumption, it becomes a necessary condition for causation [110]. At the same time, some work tried to connect time series to the real causality. In [39], Eichler had a comprehensive discussion about how to define causality for time series data. For a guide of data preparation for learning causal relationships in time series data, please refer to [93]. We can represent a time series as a chain causal graph (Fig. 9) so that algorithms for i.i.d. data can be adapted to learn causal relationships in time series.

Constraint-based Algorithms for Time Series. This class of algorithms for learning causal relationship in time series are based on the statistical independence tests. For example, FCI algorithm can be adapted [31, 40] for time series.

TiMINo. A more robust algorithm based on non-linear independent tests, known as *time series models with independent noise* (TiMINo) [109], can avoid discovering false causation with misspecified model. TiMINo takes time series data as input and either outputs a DAG or remains undecided. Given J-dimensional time series data of length L and window length p, $(\boldsymbol{x}(1), ..., \boldsymbol{x}(L))$, we start with $S = \{1, ..., J\}$. Then we describe what TiMINo does for an iteration. For each variable $x_{,j}$, we fit TiMINo for $x_{,j}(l)$ using $x_{,j}(l-1), ..., x_{,j}(l-p)$ and $x_{,k}(l), ..., x_{,k}(l-p), k \neq j$. Specifically, vector autoregression (VAR), generalized additive models (GAM) and Gaussian Process (GP) can be used for the fitting. For example, with VAR, we fit $f_j(x_{,j}(l-1), ..., x_{,j}(l-p), x_{,k}(l), ..., x_{,k}(l-p), ...) = a_j(l-1)x_{,j}(l-1) + ... + a_k(l)x_{,k}(l) + ...$ to estimate $x_{,j}$. Then we test if residuals of these models are independent of $x_{,j}, j = 1, ..., J$. Next, we choose j^* to be the variable with the weakest dependence. If there is no variable with independence, we terminate the loop and output that the causal relations remain undecided. At the end of each iteration, we remove j^* from S and set $\boldsymbol{Pa}_{j^*} = S$. We terminate the algorithm till $|S| = 1$. After removing the parents that are not required to obtain the independent residuals for each variable, we return the DAG in the form of $(\boldsymbol{Pa}_1, ..., \boldsymbol{Pa}_J)$.

Algorithms for FCMs. Those algorithms for FCMs (e.g., ICA-LiNGAM) can also be adapted to handle time series. They are also based on the asymmetry in cause-effect pairs in these models. For example, an auto-regressive LiNGAM is proposed to learn causal relationships [68].

[12]Here, the term Granger causality is also used to refer to its nonlinear variants.

With these well established algorithms, there are challenges of learning causal relationships in time series from the data perspective. Two issues can happen in the data collection process: the subsample problem and hidden time series. The subsample problem refers to the situation that only a low-resolution version of the original time series is available, for example, we only observe the time series every k time steps [133]. The key assumption for the recently proposed methods [48, 66, 113] is that there exists a *true timescale* or *causal frequency* at which we can discover proper causal graph structures with the highest confidence [34]. Furthermore, for learning causal relationship in time series data, hidden time series acts like unobserved confounders in i.i.d. data. Confounding bias can lead to faulty causal conclusions [133]. Geiger et al. [45] showed that causal relationships can be discovered under confounding bias with several assumptions.

5 CONNECTIONS TO MACHINE LEARNING

In this section, we discuss the connections between learning causality and the following machine learning problems: supervised and semi-supervised learning, domain adaptation, and reinforcement learning. For each problem, we explore two aspects: How can causal knowledge improve prediction performance? How can machine learning help answer causal questions?

5.1 Supervised Learning and Semi-supervised Learning

Supervised Learning. We can connect learning causality to supervised learning in two aspects: (1) advanced supervised learning methods can be leveraged to learn causality; (2) from a data perspective, some problems of learning causality can be reduced to supervised learning problems.

5.1.1 Advanced Supervised Learning for Learning Causal Effects. Based on the success of machine learning techniques, more advanced methods have been developed for learning causal effects. Here we cover several categories of widely used and recently proposed methods for learning causal effects: *improved traditional methods with neural networks*, *representation learning for confounders*, *learning heterogeneous causal effects with sparse models*, and *ensembles*.

Learning Causal Effects with Neural Networks. The most straightforward way to learn causal effects with neural networks is to learn representations for features. In the study of the causal effect of forming a group on receiving loan in the microfinance platform Kiva [13], GloVe [107] and Recurrent Neural Networks (RNN) [95] are used to embed text features into a low-dimensional space. Here we denote the text features by $\boldsymbol{X}_t$ and the other features by $\boldsymbol{X}_e$ and assume the vocabulary size is M. We use ReLU [98] as the activation function. In these two models, GloVe and RNN are applied to learn representations for the text features as:

$$\boldsymbol{H} = ReLU(\boldsymbol{W}_1 f(\boldsymbol{X}_t) + \boldsymbol{W}_2 h(\boldsymbol{X}_e)), \tag{37}$$

where $f(\boldsymbol{X}_t)$ is the mapping function. Then we can predict potential outcomes by fitting a function $f(\boldsymbol{h}, t)$ to infer factual outcomes y. In [111], Pham and Shen also proposed to improve the traditional methods by applying neural networks to estimate the probability distributions such as $\hat{P}(y|t, \boldsymbol{x})$ and $\hat{P}(t|\boldsymbol{x})$.

Learning Representation of Confounders. A series of recent work for learning causal effects learns representation of confounders instead of relying on observed features. The core assumption is that we can learn representations for the confounders, which are considered to be a better approximation of the confounders than the features. It allows us go beyond the strong ignorability assumption. With specific deep learning models such as the *Balancing Counterfactual Regression* [74], the *TARnet* [126], and the *Causal Effect Variational Autoencoder* (CEVAE) [89], we can

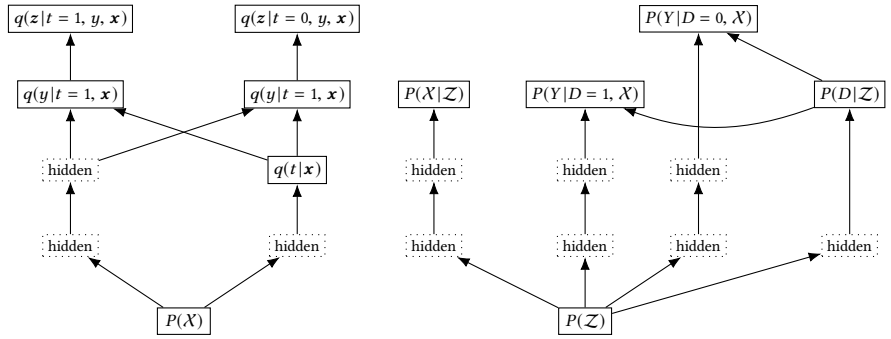
[13] https://www.kiva.org/

learn representations $\boldsymbol{z}_i$ of each instance i based on $(\boldsymbol{x}_i, d_i, y_i)$. Here, we introduce the most recent method, namely the CEVAE, which represents advances along this line.

With the recent advances in variational inference for deep latent variable models, Louizos et al. [89] proposed the CEVAE. The CEVAE consists of the inference network and the model network. The inference network is the encoder. Given an instance $(\boldsymbol{x}_i, t_i, y_i)$, the encoder learns a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_z, \Sigma_z)$ from which we can sample its latent representation $\boldsymbol{z}_i$. Then, the model network is the decoder that reconstructs the data from the latent representation. The two neural networks are shown in Fig. 10. Those variational distributions ($q(\cdot)$) approximate the corresponding infeasible posterior distributions. Similar to the VAE [76] for predictive tasks, the CEVAE is trained through minimizing the KL divergence between the data and its reconstruction. So the loss function is formulated as:

$$\mathcal{L} = \sum_i E_{q(\boldsymbol{z}_i|\boldsymbol{x}_i,t_i,y_i)}[\log P(\boldsymbol{x}_i, t_i|\boldsymbol{z}_i) + \log P(y_i|t_i, \boldsymbol{z}_i) + \log P(\boldsymbol{z}_i) - \log q(\boldsymbol{z}_i|\boldsymbol{x}_i, t_i, y_i)]. \tag{38}$$

The main difference between the CEVAE and the regular VAE is that, in CEVAE, there is a data point, $(\hat{y}_i^t, t_i, \hat{\boldsymbol{x}}_i)$ reconstructed for each combination of instance and treatment (i, t), which enables the inference of counterfactual outcomes once the neural networks in Fig. 10 are trained.



(a) The inference network (encoder). (b) The model network (decoder).

Fig. 10. The neural network structures of CEVAE. The parameters, i.e., mean and variance, of each variational distribution $q(\cdot)$, are outputs of the neural network layers below it.

Learning Heterogeneous Causal Effects with Ensembles. Ensemble models achieve the state-of-the-art performance in many supervised learning problems. With ensemble models, we train a series of weak classifiers on random subsamples of data (i.e., Bootstrapping) and make predictions by aggregating their outputs (i.e., Bagging). Variants of ensemble models are developed toward learning causal effects. In [62], Hill proposed to apply Bayesian Additive Trees (BART) [30] to estimate CATE. In particular, BART takes the features and the treatment as input and output the distribution of potential outcomes as $f(\boldsymbol{x}, t) = \mathbb{E}[y|t, \boldsymbol{x}]$, which returns the sum of the outputs of Q Bayesian regression trees as:

$$f(\boldsymbol{x}, t) = \sum_{j=1}^{Q} g_j(\boldsymbol{x}, t). \tag{39}$$

Then we can estimate the CATE for given $\boldsymbol{x}$ as $\hat{\tau}(\boldsymbol{x}) = f(\boldsymbol{x}, 1) - f(\boldsymbol{x}, 0)$. Each subtree is defined by the tree structure and a set of b leaf nodes $\{\mu_{j1}, ..., \mu_{jb}\}$. An example of a BART subtree is shown in Fig. 11, where each interior node (rectangle) sends an instance to one of its children. The k-th node of the j-th subtree has a parameter μ_{jk}, i.e., the mean outcome of the instances classified to this

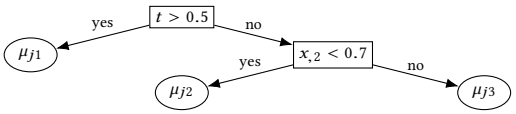


Fig. 11. A subtree $g(\boldsymbol{x}, t)$ in BART

node. BART has several advantages [57, 62]: (1) it is good at capturing non-linearity and discontinuity, (2) it frees researchers from hyperparameter tuning and (3) it outputs posterior distribution of outcomes, which allows uncertainty quantification. Hahn et al. [57] proposed to handle the problem regularization-induced confounding (RIC) with BART [56]. RIC happens when the potential outcomes heavily depend on the features rather than the treatment.

In [141], Wager and Athey proposed the Causal Forest, which can output asymptotically normal and consistent estimation of CATE. Each tree in the causal forest partitions the original covariate space recursively into subspaces such that each subspace is represented by a leaf. Function $L_j(\boldsymbol{x})$ returns which leaf of the j-th causal tree in the forest a certain instance belongs to, given its features $\boldsymbol{x}$. Then leaf of the j-th tree is considered as a RCT such that the CATE of a given $\boldsymbol{x}$ is identified and can be estimated by $\hat{\tau}_j(\boldsymbol{x}) = \frac{1}{|U_l^1|}\sum_{i\in U_l^1} Y_i - \frac{1}{|U_l^0|}\sum_{i\in U_l^0} Y_i$, where $U_l^t = \{i|t_i = t, L_j(\boldsymbol{x}_i) = l\}$ refers to the subset of instances that are sent to the l-th leaf of the j-th subtree whose treatment is t. Then the causal forest simply outputs the average of the CATE estimation from the J trees as $\hat{\tau}(\boldsymbol{x}) = \frac{1}{J}\sum_j \hat{\tau}_j(\boldsymbol{x})$. It is worth mentioning that there are studies dealing with the case where heterogeneous subpopulations cannot be identified by features such as *principle stratification* [42, 140].

Methods for Non-i.i.d. Data. In some cases, an instance's treatment or outcome can depend on those of other instances. For example, the customer flow of restaurants in the same area may amount to a constant. Besides the features, treatments, and outcomes, auxiliary information can be utilized to capture the dependence between instances. As mentioned in Section 1, such interdependency can be networks, time series, or temporal point process. Possible solutions for learning causal effects on non-i.i.d data include modeling the interference [114] and disentangle instances with i.i.d. representations. Then the methods for i.i.d. data can be used. For example, in [114], a linked VAE is developed to handle spillover effect where an instance's treatment affects its neighbors' outcomes. In general, learning causal effects from non-i.i.d. data is still an open problem.

5.1.2 Learning Causality as Supervised Learning. Now we discuss how supervised learning algorithms can help learning causality. The problem of learning causal relationships can be transformed into as a prediction problem once we label the data with causal relationships. In particular, suppose we are given labeled training data of the form $(c_1, a_1), ..., (c_N, a_N)$ where each c_j is an i.i.d. dataset $c_j = (\boldsymbol{X}_1, \boldsymbol{y}_1), ..., (\boldsymbol{X}_{N_j}, \boldsymbol{y}_{N_j})$ sampled from a joint distribution $P_j(\boldsymbol{x}, y)$ and each dataset has an additional label $a_j \in (\rightarrow, \leftarrow)$ describing whether the dataset is *causal* $\boldsymbol{x} \rightarrow y$ or *anti-causal* $y \rightarrow \boldsymbol{x}$. The main challenge here is to obtain the label of causal direction. For some datasets, the causal relationships is naturally revealed [88]. In addition, we can leverage the knowledge that a dataset is causal or anti-causal to improve supervised learning models. Causal regularization [15, 127] is proposed to learn more interpretable and generalizable models. In [15], a causal regularizer guides predictive models towards learning causal relationships between features and labels. It is assumed that a classifier $c^i = P(x^i \text{ does not cause } y)$ outputs whether a feature x^i causes the label y. Then the objective function of a predictive model with the causal regularizer is formulated as:

$$\arg\min_{\boldsymbol{w}} \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}(\boldsymbol{x}_j, y_j|\boldsymbol{w}) + \lambda \sum_{i=1}^{m} c^i|w^i|, \tag{40}$$

where $\mathcal{L}$ denotes the loss function of the predictive model with parameters $\boldsymbol{w}$. Intuitively, the lower the probability of a feature to be a cause, the more penalty will be added to its corresponding weight, which eventually encourages the model to pay more attention to those features that are more likely to be causes of the label. In [78], the following causal regularizer is proposed to set each feature as the treatment and learn sample weights such that the distribution of treated and control group can be balanced w.r.t. to each treatment (feature):

$$\sum_{j=1}^{m} || \frac{\boldsymbol{X}_{,-j}^T(\boldsymbol{w} \odot \boldsymbol{I}_{,j})}{\boldsymbol{w}^T \boldsymbol{I}_{,j}} - \frac{\boldsymbol{X}_{,-j}^T(\boldsymbol{w} \odot (\boldsymbol{e} - \boldsymbol{I}_{,j}))}{\boldsymbol{w}^T(\boldsymbol{e} - \boldsymbol{I}_{,j})} ||_2^2, \tag{41}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ signifies the sample weights, $\boldsymbol{e}$ denotes the $n \times 1$ vector with all elements equal to 1, $\boldsymbol{X}_{,j}$ and $\boldsymbol{X}_{,-j}$ are the j-th column of the feature matrix and the matrix of remaining features, $\boldsymbol{I}_{i,j}$ refers to the treatment status of the i-th instance when the j-th feature is set as the treatment. The authors added a constraint to the original loss function of a logistic regression model to ensure the value of this causal regularizer is not greater than a predefined hyperparameter $\gamma \in \mathbb{R}^+$. The authors claimed that doing this can help identify causal features and construct robust predictive model across different domains.

Semi-supervised Learning (SSL). A machine learning problem can be either causal or *anti-causal* [122]. Anti-causal means that the label y is the cause of the features $\boldsymbol{x}$. For example, in hand written digits recognition [83], which digit to write is first determined, then the digit would be represented as a matrix of pixel values. Such an causal structure has implications for SSL. In SSL, the target is to improve the quality of estimated $P(y|\boldsymbol{x})$ with additional unlabeled instances which can provide information of the marginal distribution $P(\boldsymbol{x})$. We can first consider the cases when semi-supervised learning would fail. For example, when $p(\boldsymbol{x})$ is a uniform distribution, observing more unlabeled instances provide no information about $P(y|\boldsymbol{x})$. In contrast, in the case of representation learning, if $\boldsymbol{h}$ contains hidden causes of the features $\boldsymbol{x}$, and the label y is one of the causal factors of $\boldsymbol{x}$, then predicting y from $\boldsymbol{h}$ is likely to be easy [50]. Specifically, the true data-generating process implies $\boldsymbol{h}$ is a parent of $\boldsymbol{x}$, and thus, $p(\boldsymbol{h}, \boldsymbol{x}) = P(\boldsymbol{x})P(\boldsymbol{x}|\boldsymbol{h})$. So the marginal distribution is $P(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{h}}[P(\boldsymbol{x}|\boldsymbol{h})]$. With $P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{p(\boldsymbol{x})}$, we know $P(\boldsymbol{x})$ directly affects $P(y|\boldsymbol{x})$. Therefore, the causal structure of $P(\boldsymbol{x})$ can help the prediction of $P(y|\boldsymbol{x})$. However, the number of causes can be extremely large. For example, a positive Yelp review can result from good service, delicious food, cheap price, or decent restaurant environment. Brute force solutions are not feasible as it is often impossible to capture most of the causes. Therefore, we need to figure out what causes to encode for a certain target y. Fixed criteria such as mean squared error on reconstructed features have been used to train autoencoders and generative models, which assumes that a latent variable is salient iff it affects the value of most features. However, there can be tasks where the label is only associated with few causes. For example, to predict the customer flow of truck drivers in fast food restaurants near highway, few hidden causes may be useful. Therefore, the criteria need to be adaptable in accordance with the task. Generative Adversarial Networks (GAN) [51] are proposed to address this issue for images. GAN can adapt its criteria s.t. the latent variables (e.g., ears of human in human head images) that only affect the value of few features can also be learned as representations of data. Finding optimal ways to decide which causes we learn representation for is still an open problem.

Janzing and Schölkopf [72] considered a special case of SSL: let $y = f(x)$ and $x, y \in [0, 1]$, where f is an unknown anti-causal model. Given $n-1$ labeled instances $\{(x_i, y_i)\}_{i=1}^{n-1}$, we seek to infer the label $y_n = f(x_n)$ of an unlabeled instance x_n. In this setting, it is proved that SSL outperforms supervised learning when $P(x)$ and f are dependent and a certain independence between $P(y)$ and $g = f^{-1}$ holds. The independence $P(y) \perp\!\!\!\perp g$ is assumed and can be defined as $Cov[P(y), \log g'] = 0$,

where g' denotes the derivative of g and $\log g', P(x) \in [0,1]$. Given that, it can be shown that $Cov[P(x), \log f'] > 0$, which means $P(x)$ contains information of the function f we aim to learn. In addition, the fact that SSL only works in the anti-causal direction can help learn causal relationships [125]. As shown above, if the problem is anti-causal, we expect that better knowledge of $P(x)$ helps prediction of $P(y|x)$ as they are dependent. In contrast, if it is a causal problem, then knowing $P(x)$ barely helps us learn $P(y|x)$. Therefore, comparing the errors of estimations on $P(y|x)$ and $P(x|y)$ enables us to determine the direction of causality. In particular, Gaussian process (GP) regression models are trained to estimate $P(x|y)$ as:

$$P(x|y, \boldsymbol{y}^*) = \int_{\mathcal{Z},\theta} P(\boldsymbol{z}, \boldsymbol{\theta}, x|\boldsymbol{y}^*, y) d\boldsymbol{z} d\boldsymbol{\theta} \approx \int_{\mathcal{Z},\theta} P(x|y, \boldsymbol{y}^*, \boldsymbol{z}, \boldsymbol{\theta}) P(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y}^*) d\boldsymbol{z} d\boldsymbol{\theta}, \tag{42}$$

where $\boldsymbol{z}$ signifies the latent variables and $\boldsymbol{y}^* = (y_1, ..., y_{n-1})$ are the observed data points. $\mathcal{Z}$ and θ are the set of possible values of the latent variables and the model parameters, respectively. The first factor $P(x|y, \boldsymbol{y}^*, \boldsymbol{z}, \boldsymbol{\theta})$ is the supervised GP regression and the second factor $P(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y}^*)$ denotes the posterior distribution over $\boldsymbol{z}$ and $\boldsymbol{\theta}$ given observed labels $\boldsymbol{y}^*$. Assuming uniform priors for $\boldsymbol{z}$ and $\boldsymbol{\theta}$, using Bayes's rule $p(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y}^*) = \frac{p(\boldsymbol{y}^*|\boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z}) p(\boldsymbol{\theta})}{p(\boldsymbol{y}^*)} \propto p(\boldsymbol{y}^*|\boldsymbol{z}, \boldsymbol{\theta})$ which is parameterized by a Gaussian distribution defined by GP-LVM [137]. Thus, we can estimate $P(x|y, \boldsymbol{y}^*) = \frac{1}{m} \sum_i p(x|y, \boldsymbol{y}^*, \boldsymbol{z}^i, \boldsymbol{\theta}^i)$ with m MCMC samples from $p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}^*)$. In a similar way, we can find that $p(x|y, \boldsymbol{y}^*, \boldsymbol{z}^i, \boldsymbol{\theta}^i)$ is also proportional to a Gaussian distribution defined by GP-LVM. Thus, we can estimate $P(x|y)$ as mentioned above and $P(y|x)$ by repeating the procedure with x and y swapped. Finally, we compare the log likelihood of the two estimations to figure out the causal direction. However, it is still an open problem to scale this type of approaches (regression for causal discovery) to high-dimensional, noisy and non-i.i.d. data.

5.2 Domain Adaptation

Domain adaptation [21, 35] studies how to adapt machine learning models trained in some domains to the others. One real-world application of domain adaptation is to improve prediction accuracy when we have plenty of labeled data for the source domain (e.g., Yelp review) but not for the target domain (e.g., Tripadvisor[14] review). Domain adaptation is naturally related to learning causality by *invariant prediction in different domains* [108]. Suppose we have observed data with a target variable y^e and m predictor variables $\boldsymbol{x}^e = (x_1^e, ..., x_m^e)$ from different domains $e \in \{1, ..., E\}$ and the target is to predict the value of y. Invariant prediction assumes that the conditional $P(y|\boldsymbol{Pa}_y)$ is consistent for all domains, where $\boldsymbol{Pa}_y$ is a set of direct causes of y. Formally,

$$P(y^e|\boldsymbol{Pa}_y^e) = P(y^f|\boldsymbol{Pa}_y^f). \tag{43}$$

The assumption is valid when the distributions are induced by an underlying SCM and the different domains correspond to different intervention distributions, for which y has not been intervened on. Then we can conclude that (1) invariant prediction is achieved and (2) $\mathcal{Z}^*$ is the set of estimated causes of the target variable y. In [108], the authors proposed a method to estimate $\boldsymbol{Pa}_y$. Assuming that the collection $\mathcal{S}$ consists of all subsets S of features that result in *invariant prediction*, satisfying $P(y^e|x_S^e) = P(y^f|x_S^f)$, the variables appearing in each such set S form the estimated causes of the label $\boldsymbol{Pa}_y$. Finally, a valid subset S^* that achieves the best performance in the source domains is selected as the features for cross domain prediction. This is because the selected subset is guaranteed to be optimal in terms of domain generalization error. Due to the independent mechanism assumption [110, 118], the selected subset is also robust against arbitrary changes of marginal

[14] https://www.tripadvisor.com/

marginal distribution of predictors in the target domains. Similar results for domain generalization have been obtained through a global balancing approach [78] and a causal feature selection method [101]. They are based on sample re-weighting. In [78], the proposed model, Deep Global Balancing Regression (DGBR), leverages a auto-encoder model to map data into a latent space before reweighting instead of directly reweighting the original samples [101]. We summarize the usage of the low-dimensional representations of DGBR in two ways: (1) They are used in the global balancing regularizer, where each variable is successively set as the treatment variable. Then we balance all the variables via learning global sample weights. (2) We can predict outcomes based on the representations using regularized regression. The causal regularizer of DGBR is:

$$\sum_{j=1}^{p} \left\| \frac{\phi(\boldsymbol{X}_{,-j})^T(\boldsymbol{w} \odot \boldsymbol{X}_{,j})}{\boldsymbol{w}^T \boldsymbol{X}_{,j}} - \frac{\phi(\boldsymbol{X}_{,-j})^T(\boldsymbol{w} \odot (\boldsymbol{e} - \boldsymbol{X}_{,j}))}{\boldsymbol{w}^T(\boldsymbol{e} - \boldsymbol{X}_{,j})} \right\|_2^2, \tag{44}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ signifies the sample weights, $\boldsymbol{e}$ denotes the $n \times 1$ vector with all elements equal to 1, $\boldsymbol{X}_{,j}$ and $\boldsymbol{X}_{,-j}$ are the j-th column of the feature matrix and the matrix of remaining features. In addition, the objective function is a weighted loss of logistic regression along with a constraint to limit the value of this causal regularizer not greater than a predefined positive hyperparameter. While for prediction under concept drift [143], where Eq. 43 is violated but the marginal distribution of predictors remain, one may allow apriori causal knowledge to guide the learning process and circumvent the discrepancies between the source and target domains [106], a.k.a *causal transportability*. The study of transportability seeks to identify conditions under which causal information learned from experiments can be reused in different domains. A formal definition of causal transportability can be referred to [106]. In [19], the authors further provide a necessary and sufficient condition to decide, given assumptions about differences between the source and target domains, whether transportability is feasible.

5.3 Reinforcement Learning

Reinforcement learning (RL) [134] is studied for solving sequential decision-making problems. The key variables in RL are the action a, the state z, and the reward y. When an agent performs an action, it reaches the next state and receives a reward. The *Markov decision process* is often adopted, where the next state z_{t+1} depends on the current state z_t and action a_t and the reward of the next state y_{t+1} is determined by z_t, z_{t+1} and a_t. A RL model learns a *policy* $\pi(a_t, z_t) = P(a_t|z_t)$ which determines which action to take given the current state. The objective is to maximize the sum of the rewards. In the running example, we can assume that the state z_t represents the location of a restaurant, the action a_t can be moving to a certain place or staying at the same place and the reward is the customer flow y. In each time step, the restaurant owner decides which action to take and then observe the customer flow. Then the owner will make decisions for the next time step based on whether the customer flow increases or not.

Unobserved Confounders in RL. Unobserved confounders raise issues of learning policies for RL models such as multi-armed bandits (MAB) [18]. Without knowing the causal model, MAB algorithms can perform as badly as randomly taking an action in each time step. Specifically, the Causal Thompson Sampling algorithm [18] is proposed to handle unobserved confounders in MAB problems. The reward distributions of the arms that are not preferred by the current policy can also be estimated through hypothetical interventions on the action (choice of arm). By doing this we can avoid confounding bias in estimating the causal effect of choosing an arm on the expected reward. To connect causality with RL, we view a strategy or a policy in RL as an intervention [110].

Unbiased Reward Prediction. Given trajectories (actions, states and rewards) of an observed strategy, we can utilize causal inference methods to predict rewards for another strategy, especially for Episodic RL (ERL) problems. ERL is a subclass of RL where the state is reset to the default value after a finite number of actions. ERL helps decision-making in many applications such as card games and advertisement placement [22]. One popular approach leverages IPTW for predicting reward of ERL models. In IPTW, a treatment refers to an action and the strategy-specific propensity score is defined as the probability to perform the selected action given the observed state. Particularly, given trajectories produced by running an observed strategy π L times $[(a_1(1), z_1(1)), (a_2(1), z_2(1)), ...], ..., [(a_1(n), z_1(n)), (a_2(n), z_2(n)), ...]$, we can estimate the expected sum of rewards of a strategy $\tilde{\pi}$ with IPTW as:

$$\hat{\xi} := \frac{1}{L}\sum_{l=1}^{L} y(l)\frac{\prod_{k=1}^{K}\tilde{\pi}(a_k(l)|z_k(l))}{\prod_{k=1}^{K}\pi(a_k(l)|z_k(l))}, \tag{45}$$

where K denotes the number of time steps in each episode. In [22, 135], improved variants of this approach are proposed.

RL with Auxiliary Causal Knowledge. There is a line of work to improve RL models with causal knowledge as side information [81, 144]. Here, we use the Causal Bandit (CB) problem [81] as an example. In this problem, given J binary variables $x_{,1}, ..., x_{,J}$ and their causal graph but not the causal mechanisms $x_{,j} = f(\boldsymbol{Pa}_j, \epsilon_j)$, we aim to find the intervention that is most likely to set a specified variable $x_{,k}$ to 1. An intervention is defined as a vector $\boldsymbol{a} \in \{*, 0, 1\}^J$, where $a_j \neq *$ means $x_{,j}$ is set to a_j. For exploration, a CB algorithm learns $\mu(\boldsymbol{a}) = P(x_{,k} = 1|do(\boldsymbol{a}))$, the probability that the target $x_{,k}$ is set to 1 by $\boldsymbol{a}$. For exploitation, a CB algorithm minimizes the regret $R = \mu(\boldsymbol{a}^*) - \mathbb{E}[\mu(\hat{\boldsymbol{a}})]$, where $\boldsymbol{a}^*$ is the optimal action and $\hat{\boldsymbol{a}}$ denotes the algorithm's selection. In [81], the parallel bandit (PB) algorithm is proposed to solve this problem with guarantee to outperform non-causal MAB algorithms. Given total rounds L, in first $L/2$ rounds, the PB algorithm collects observational data by doing intervention $\boldsymbol{a} = [*, ..., *]$. Then it analyzes the observational data for each intervention $\boldsymbol{a} = do(x_{,j} = x)$ to estimate the reward as $\hat{\mu}(\boldsymbol{a}) = \frac{1}{L_{\boldsymbol{a}}}\sum_{l=1}^{L/2}\mathbb{1}(x_{l,j} = x_{,j}(\boldsymbol{a}))$ and probabilities as $\hat{p}_{\boldsymbol{a}} = \frac{2L_{\boldsymbol{a}}}{L}$, $\hat{q}_j = \hat{p}_{do(x_{,j}=1)}$, where $L_{\boldsymbol{a}} = \sum_{l=1}^{L/2}\mathbb{1}(x_{l,j} = x)$ denotes the number of times we observe what $\boldsymbol{a}$ could have done in the observational data. Next, we create the set of rarely observed actions as $\mathcal{A}' = \{\boldsymbol{a}|\hat{p}_{\boldsymbol{a}} \geq \frac{1}{\hat{m}}\}$, where $\hat{m} =$. Then we uniformly sample $\boldsymbol{a} \in \mathcal{A}'$ and observe the value of $x_{,k}$. At the end, we compute $\mathbb{E}[x_{,k}]$ of resulting each action as the estimated reward $\hat{\mu}(\boldsymbol{a})$ and select the one with the largest $\hat{\mu}(\boldsymbol{a})$. Other work bridging RL and causality includes causal approaches for transfer learning in RL models [145] and data-fusion for reinforcement learners [41].

At the end, we summarize advantages and disadvantages of causal machine learning. The advantages of machine learning with causality include: (1) invariant prediction under environment changes [78, 101, 108], (2) model generality and interpretability [15, 127], and (3) performance improvement with theoretical guarantee [81, 144]. On the other hand, causal machine learning mainly faces the challenges of insufficient amount of data. Causal machine learning algorithms may require ground truth of counterfactuals [74, 126] or interventional data [81, 144] for training or evaluation, which can be difficult to collect.

6 CONCLUSIONS AND SOME OPEN PROBLEMS

Different from previous surveys, this work aims to solve the problem of learning causality under the big data setting where we have more data and less knowledge than the traditional causal studies. Although the methods in the existing literature may not directly address learning causality for such cases, they build the foundation of data-driven approaches for both learning causal effects and

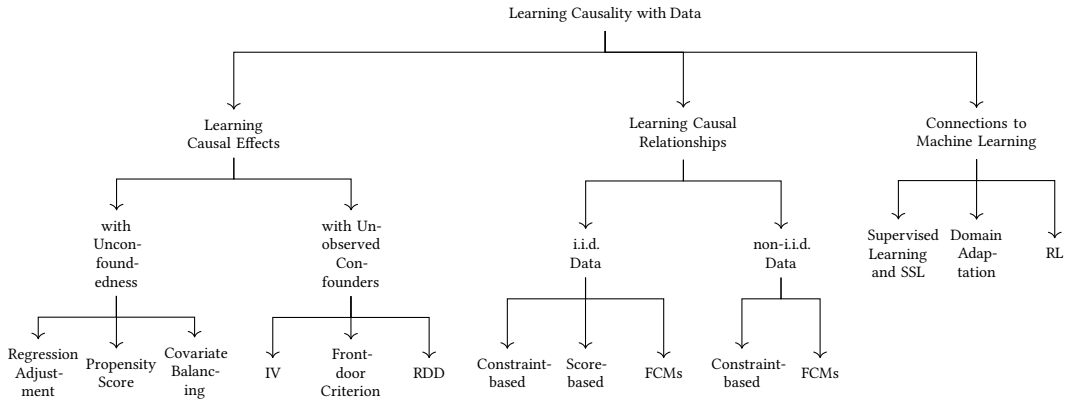


Fig. 12. Learning causality with data: a summary of the survey

relationships. Another idea highlighted in this work is the connections between learning causality and machine learning. We aim to demonstrate that it is possible to leverage the connections between them in achieving better solutions for both causal and predictive problems. Moreover, machine learning models can benefit from exploiting learned causal knowledge in Section 5.

Fig. 12 shows a summary of all the contents covered in this survey. These two frameworks enable us to formulate problems of learning causality with mathematical languages. Then, we cover the two types of problems: learning causal effect (causal inference) and relationships (causal discovery) with data. The methods for learning causal effects with three types of data are presented: i.i.d. and non-i.i.d. data with back-door criterion satisfied and data with unobserved confounders. Next, we discuss how to learn causal relationships from two types of data: i.i.d. data and time series data. Finally, we discuss the connections between learning causality and machine learning. In particular, we discuss how we can help learning causality with methods for solving the three families of machine learning problems: supervised and semi-supervised learning, domain adaptation and reinforcement learning. We describe the connections from two perspectives: how learning causality yields better prediction in the machine learning problem? How machine learning techniques can be applied for learning causality?

Even existing research solved some problems of learning causality with data, more work needs to be done toward addressing the challenges that come with big data. From the data perspective, we present some open problems to review the great potential of learning causality with data:

- Study of heterogeneous groups: A dataset can come with heterogeneous groups. Though existing work addressed this problem by showing difference between groups [3] in terms of causal effects. But more efficient methods are needed for massive data. An extreme case of heterogeneous groups are anomalies. Although anomaly detection has been well studied as a prediction problem [1, 2, 26], detection of instances that are irregular in causal effects and relationships is still an open problem.
- Learning causality with imbalanced data: For example, a dataset for learning causal effect can come with very few treated units but much more controlled units. The problem of learning causality for such data remains to be solved.
- Learning causality with complex variables: An example of this type is the problem of learning causal effect of taking courses on employment. The treatment variable, namely the courses can be taken, can be formulated as a knowledge graph. So the problem turns into

learning the causal effect of knowledge graphs on employment. The complex variable can also be the outcome or other variables and multiple complex variables can appear in the same problem.

APPENDIX

To facilitate development, evaluation and comparison of methods for learning causality, we introduce the open source data and algorithm index.

Data Index for Learning Causality We develop the open source data index for learning causality. It is available at Github (https://github.com/rguo12/awesome-causality-data). The datasets are categorized by the problem and the type of data.

Algorithm Index for Learning Causality The open source algorithm index for learning causality is at Github (https://github.com/rguo12/awesome-causality-algorithms). This index lists the algorithms mentioned in this survey. We group the algorithms by the problem and the type of data.

REFERENCES

[1] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. 2010. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*. 410–421.

[2] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *DMKD* 29, 3 (2015), 626–688.

[3] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data. *arXiv preprint arXiv:1805.03094* (2018).

[4] Dionissi Aliprantis. 2015. A distinction between causal effects in Structural and Rubin Causal Models. (2015).

[5] Yakov Amihud. 2002. Illiquidity and stock returns: cross-section and time-series effects. *JFM* 5, 1 (2002), 31–56.

[6] Michael Anderson and Jeremy Magruder. 2012. Learning from the Cloud: Regression Discontinuity Estimates of the Effects of an Online Review Database. *EJ* 122, October (2012), 957–989.

[7] Joshua D Angrist and Guido W Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* 90, 430 (1995), 431–442.

[8] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455.

[9] Joshua D Angrist and Victor Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.* 114, 2 (1999), 533–575.

[10] Sinan Aral and Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature communications* 8 (2017), 14753.

[11] David Arbour, Dan Garant, and David Jensen. 2016. Inferring Network Effects from Observational Data. In *KDD*. 715–724.

[12] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015).

[13] Susan Athey, Guido W Imbens, and Stefan Wager. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Series B (Stat. Methodol.)* 80, 4 (2018), 597–623.

[14] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46, 3 (2011), 399–424.

[15] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. 2017. Causal regularization. *arXiv preprint arXiv:1702.02604* (2017).

[16] Mohammad Taha Bahadori and Yan Liu. 2012. Granger causality analysis in irregular time series. In *SDM*. 660–671.

[17] Mohammad Taha Bahadori and Yan Liu. 2013. An examination of practical granger causality inference. In *SDM*. 467–475.

[18] Elias Bareinboim, Andrew Forney, and Judea Pearl. 2015. Bandits with unobserved confounders: A causal approach. In *NIPS*. 1342–1350.

[19] Elias Bareinboim and Judea Pearl. 2012. Transportability of causal effects: Completeness results. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[20] Elias Bareinboim and Jin Tian. 2015. Recovering Causal Effects from Selection Bias.. In *AAAI*. 3475–3481.

[21] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP*. 120–128.

[22] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR* 14, 1 (2013), 3207–3260.
[23] Donald T Campbell. 1969. Reforms as experiments. *Am. Psychol.* 24, 4 (1969), 409.
[24] Christopher Carpenter and Carlos Dobkin. 2009. The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *AEJ: Applied Economics* 1, 1 (2009), 164–82.
[25] Nancy Cartwright et al. 1994. Nature's Capacities and their Measurement. *OUP Catalogue* (1994).
[26] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41, 3 (2009), 15.
[27] David Maxwell Chickering. 1996. Learning Bayesian networks is NP-complete. In *Learning from data*. Springer, 121–130.
[28] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.
[29] David M Chickering, Dan Geiger, David Heckerman, et al. 1994. *Learning Bayesian networks is NP-hard*. Technical Report. Citeseer.
[30] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4, 1 (2010), 266–298.
[31] Tianjiao Chu and Clark Glymour. 2008. Search for additive nonlinear time series causal models. *JMLR* 9, May (2008), 967–991.
[32] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* (2012), 294–321.
[33] Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.
[34] David Danks and Sergey Plis. 2013. Learning causal structure from undersampled time series. (2013).
[35] Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).
[36] Rajeev H Dehejia and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* 94, 448 (1999), 1053–1062.
[37] Imme Ebert-Uphoff and Yi Deng. 2012. Causal discovery for climate research using graphical models. *JCLI* 25, 17 (2012), 5648–5665.
[38] Andrew C Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. 2018. Regression discontinuity designs based on population thresholds: Pitfalls and solutions. *Am. J. Pol. Sci.* 62, 1 (2018), 210–229.
[39] Michael Eichler. 2012. Causal inference in time series analysis. *Causality: Statistical perspectives and applications* (2012), 327–354.
[40] Doris Entner and Patrik O Hoyer. 2010. On causal discovery from time series data using FCI. *PGM* (2010), 121–128.
[41] Andrew Forney, Judea Pearl, and Elias Bareinboim. 2017. Counterfactual Data-Fusion for Online Reinforcement Learners. In *ICML*. 1156–1164.
[42] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.
[43] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2008. Kernel measures of conditional dependence. In *NIPS*. 489–496.
[44] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *Am. J. Epidemiol.* 173, 7 (2011), 761–767.
[45] Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *ICML*. 1917–1925.
[46] Andrew Gelman. 2011. Causality and statistical learning. *AJS* 117, 3 (2011), 955–966.
[47] Andrew Gelman and Guido Imbens. 2018. Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econ. Stat.* (2018), 1–10.
[48] Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. 2015. Discovering temporal causal relations from subsampled data. In *ICML*. 1898–1906.
[49] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. 2017. Causal discovery from temporally aggregated time series. In *UAI*, Vol. 2017.
[50] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
[51] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.
[52] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* (1969), 424–438.
[53] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*. 6645–6649.

[54] Xing Sam Gu and Paul R Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph Stat.* 2, 4 (1993), 405–420.

[55] Ruocheng Guo, Jundong Li, and Huan Liu. 2018. INITIATOR: Noise-contrastive Estimation for Marked Temporal Point Process.. In *IJCAI*. 2191–2197.

[56] P Richard Hahn, Carlos M Carvalho, David Puelz, Jingyu He, et al. 2018. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* 13, 1 (2018), 163–182.

[57] P Richard Hahn, Jared S Murray, and Carlos Carvalho. 2017. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv preprint arXiv:1706.09523* (2017).

[58] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20, 1 (2012), 25–46.

[59] David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20, 3 (1995), 197–243.

[60] David Heckerman, Christopher Meek, and Gregory Cooper. 2006. A Bayesian approach to causal discovery. In *Innovations in Machine Learning*. Springer, 1–28.

[61] Miguel Ángel Hernán, Babette Brumback, and James M Robins. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* (2000), 561–570.

[62] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph Stat.* 20, 1 (2011), 217–240.

[63] Keisuke Hirano, Guido W Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 4 (2003), 1161–1189.

[64] Patrik O Hoyer, Aapo Hyvarinen, Richard Scheines, Peter L Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. 2012. Causal discovery of linear acyclic models with arbitrary distributions. *arXiv preprint arXiv:1206.3260* (2012).

[65] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2009. Nonlinear causal discovery with additive noise models. In *NIPS*. 689–696.

[66] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. 2016. Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *PGM*. 216–227.

[67] Aapo Hyvärinen and Erkki Oja. 2000. Independent component analysis: algorithms and applications. *Neural networks* 13, 4-5 (2000), 411–430.

[68] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. 2010. Estimation of a structural vector autoregression model using non-gaussianity. *JMLR* 11, May (2010), 1709–1731.

[69] Kosuke Imai and Marc Ratkovic. 2014. Covariate balancing propensity score. *J. R. Stat. Soc. Series B (Stat. Methodol.* 76, 1 (2014), 243–263.

[70] Kosuke Imai, Marc Ratkovic, et al. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7, 1 (2013), 443–470.

[71] Guido W Imbens. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* 86, 1 (2004), 4–29.

[72] Dominik Janzing and Bernhard Schölkopf. 2015. Semi-supervised interpolation in an anticausal learning scenario. *JMLR* 16 (2015), 1923–1948.

[73] Marshall M Joffe, Thomas R Ten Have, Harold I Feldman, and Stephen E Kimmel. 2004. Model selection, confounder control, and marginal structural models: review and new applications. *Am. Stat.* 58, 4 (2004), 272–279.

[74] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.

[75] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8, Mar (2007), 613–636.

[76] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[77] Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. 2017. Cost-optimal learning of causal graphs. In *ICML*. JMLR. org, 1875–1884.

[78] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable Prediction across Unknown Environments. In *KDD*. 1617–1626.

[79] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In *KDD*. 265–274.

[80] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* (1986), 604–620.

[81] Finnian Lattimore, Tor Lattimore, and Mark D Reid. 2016. Causal bandits: Learning good interventions via causal inference. In *NIPS*. 1181–1189.

[82] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, and Huawen Liu. 2015. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *arXiv preprint arXiv:1502.02454* (2015).

[83] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. 1990. Handwritten digit recognition with a back-propagation network. In *NIPS*. 396–404.

[84] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 94.

[85] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed network embedding for learning in a dynamic environment. In *CIKM*. 387–396.

[86] Jundong Li, Osmar R Zaïane, and Alvaro Osornio-Vargas. 2014. Discovering statistically significant co-location rules in datasets with extended spatial objects. In *DaWaK*. 124–135.

[87] Qi Li and Jeffrey Scott Racine. 2007. *Nonparametric econometrics: theory and practice*. Princeton University Press.

[88] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. 2017. Discovering causal signals in images. In *CVPR 2017*.

[89] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *NIPS*. 6446–6456.

[90] Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. 2009. Grouped graphical Granger modeling methods for temporal causal modeling. In *KDD*. 577–586.

[91] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 23, 19 (2004), 2937–2960.

[92] HernÃąn MA and Robins JM. forthcoming. *Causal Inference*. CRC Boca Raton, FL:.

[93] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.

[94] Subramani Mani and Gregory F Cooper. 2000. Causal discovery from medical textual data.. In *Proceedings of the AMIA Symposium*. 542.

[95] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.

[96] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR* 17, 1 (2016), 1103–1204.

[97] Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

[98] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*. 807–814.

[99] Jersey Neyman. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10 (1923), 1–51.

[100] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. 2013. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* 381, 9875 (2013), 1371–1379.

[101] Michael J Paul. 2017. Feature Selection as Causal Inference: Experiments with Text Classification. In *CoNLL*. 163–172.

[102] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[103] Judea Pearl. 2009. Causal inference in statistics: An overview *. *Stat. Surv.* 3 (2009), 96–146.

[104] Judea Pearl. 2009. *Causality*. Cambridge university press.

[105] Judea Pearl. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* (2018).

[106] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[107] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[108] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Series B Stat. Methodol.* 78, 5 (2016), 947–1012.

[109] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. In *NIPS*. 154–162.

[110] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. MIT press.

[111] Thai T Pham and Yuanyuan Shen. 2017. A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform. *arXiv preprint arXiv:1706.02795* (2017).

[112] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. 2014. Seeing the arrow of time. In *CVPR*. 2035–2042.

[113] Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. 2015. Rate-agnostic (causal) structure learning. In *NIPS*. 3303–3311.

[114] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked Causal Variational Autoencoder for Inferring Paired Spillover Effects. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1679–1682.

[115] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* 3, 2 (2017), 121–129.

[116] Joseph D Ramsey. 2014. A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv preprint arXiv:1401.5031* (2014).

[117] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology.

[118] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 2015. A Causal Perspective on Domain Adaptation. *stat* 1050 (2015), 19.

[119] Teemu Roos, Tomi Silander, Petri Kontkanen, and Petri Myllymaki. 2008. Bayesian network structure learning using factorized NML universal models. In *ITA Workshop, 2008*. 272–276.

[120] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[121] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5 (1974), 688.

[122] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471* (2012).

[123] Gideon Schwarz et al. 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 2 (1978), 461–464.

[124] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* (2013), 2263–2291.

[125] Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. 2015. Inference of cause and effect with unsupervised inverse regression. In *AISTATS*. 847–855.

[126] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. 3076–3085.

[127] Zheyan Shen, Peng Cui, Kun Kuang, and Bo Li. 2017. On Image Classification: Correlation vs Causality. *arXiv preprint arXiv:1708.06656* (2017).

[128] Shohei Shimizu. 2014. LiNGAM: non-Gaussian methods for estimating causal structures. *Behaviormetrika* 41, 1 (2014), 65–98.

[129] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *JMLR* 7, Oct (2006), 2003–2030.

[130] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *JMLR* 12, Apr (2011), 1225–1248.

[131] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. 2000. *Causation, prediction, and search*. MIT press.

[132] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *UAI*. 499–506.

[133] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, Vol. 3. 3.

[134] Richard S Sutton and Andrew G Barto. 1998. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

[135] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*. 814–823.

[136] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* (1996), 267–288.

[137] Michalis Titsias and Neil D Lawrence. 2010. Bayesian Gaussian process latent variable model. In *AISTATS*. 844–851.

[138] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoldi. 2018. Propensity score methodology in the presence of network entanglement between treatments *. *arXiv preprint arXiv:1801.07310* (2018).

[139] Mark J Van Der Laan and Daniel Rubin. 2006. Targeted maximum likelihood learning. *Int. J. Biostat.* 2, 1 (2006).

[140] Tyler J VanderWeele. 2011. Principal stratification–uses and limitations. *Int. J. Biostat.* 7, 1 (2011), 1–14.

[141] Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* just-accepted (2017).

[142] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. 2017. Permutation-based causal inference algorithms with interventions. In *NIPS*. 5822–5831.
[143] Gerhard Widmer and Miroslav Kubat. 1996. Learning in the presence of concept drift and hidden contexts. *Machine learning* 23, 1 (1996), 69–101.
[144] Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. 2018. Causal bandits with propagating inference. *arXiv preprint arXiv:1806.02252* (2018).
[145] Junzhe Zhang and Elias Bareinboim. 2017. Transfer learning in multi-armed bandit: a causal approach. In *AMMAS*. 1778–1780.
[146] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *UAI*. 647–655.
[147] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775* (2012).
[148] Dawei Zhou, Jingrui He, Hongxia Yang, and Wei Fan. 2018. SPARC: Self-Paced Network Representation for Few-Shot Rare Category Characterization. In *KDD*. 2807–2816.

Received ; revised ; accepted