A Primer on Causal Analysis

Finnian Lattimore and Cheng Soon Ong

5 June 2018

Abstract

We provide a conceptual map to navigate causal analysis problems. Focusing on the case of discrete random variables, we consider the case of causal effect estimation from observational data. The presented approaches apply also to continuous variables, but the issue of estimation becomes more complex. We then introduce the four schools of thought for causal analysis [1].

---

[1]Parts of this document are copied verbatim from Lattimore [2018].

1 Conceptual Map

Causal inference is an intuitively seductive phrase, and its use is often clouded in mystery. This document provides a brief primer about the different kinds of problems that can be considered under the umbrella of causal inference or causal analysis. Causal inference is often contrasted with statistical or probabilistic inference, as captured by the phrase "correlation does not imply causation". Vaguely speaking causal inference is the study of adding an extra requirement to the definition of conditional probability models, such that we can mathematically express the idea of a cause. For standard statistical estimation (without an additional causal assumption), there are already challenges in estimating conditional independence from data [Agresti, 2002, Zhang et al., 2018]. These challenges do not disappear when considering causal models. It is worthwhile to also consider how we are planning to use the results of causal analysis. According to Pearl and Mackenzie [2018] there are three levels on the ladder of causation.

association Seeing and observing the environment.
  Is the incidence of lung cancer higher among smokers?

intervention Doing and intervening in the environment.
  How do we reduce lung cancer? What is the effect if we ban cigarettes?

counterfactuals Imagining, restrospection, understanding the environment.
  What if I had not smoked for the last two years?

Before we introduce the four main mathematical models of causal analysis in Section 2, we briefly discuss two other major dimensions for categorising causal inference. The first dimension (Section 1.1) distinguishes between inferring of the graph of causes and using the graph of causes to infer the effect of an action. An analogous problem exists also for non-causal statistical analysis, which distinguishes between the choice of probabilistic model class and the parameters of the model. The second dimension, which is unique to causal analysis, is the kind of data that we have available (Section 1.2). We may wish to perform causal inference based on observational data or data we obtain after performing interventions.

Given the challenges of performing causal inference, it is worth pointing out that in some sense causality is "inevitable" if we find two dependent random variables.

Reichenbachs common cause principle If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a third variable Z that causally influences both [Peters et al., 2017]. As a special case, Z may coincide with either X or Y. Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp\!\!\!\perp Y|Z$. In practice, several other reasons can give rise to dependence: There may be unobserved causes, also known as selection bias; the discovery of dependence from data might be false due to performing multiple hypothesis testing; there may be a common time dependence, for example both variables are growing exponentially.

1.1 Structure discovery vs effect estimation

Analogous to the distinction between the design choice of a class of probabilistic models and the computation of their maximum likelihood parameters, there are two types of causal analysis problems. First we may wish to discover the graph of causes from a set of data. Second given the graph, we may wish to the effect of intervening on a node in the graph. Another way of stating the distinction is to consider whether we are thinking about forward or reverse causal effects [Gelman and Imbens, 2013]. The forward direction considers the problem of estimating the effect of a given intervention, and is akin to estimating the causal strength of an edge in the graph. The reverse direction considers the problem of estimating the cause of a particular observation. The reverse direction causal analysis is similar to causal graph discovery as one would need to estimate which other random variable (from the set of all random variables) caused the effect.

1.1.1 Discovering causal graphs

Causal discovery is the attempt to learn the structure of the causal relationships between variables, on which there is a rich literature, see Spirtes and Zhang [2016] for a recent review. We may be interested in discovering a causal graph from a set of known models. This arises naturally if we assume free access to a large observational data set, from which the Markov equivalence class can be found via causal discovery techniques. Work on the problem of selecting experiments to discover the correct causal graph from within a Markov equivalence class [Eberhardt et al., 2005, Eberhardt, 2010, Hauser and Bühlmann, 2014, Hu et al., 2014] could potentially be incorporated into a causal bandit algorithm. In particular, Hu et al. [2014] show that only $\mathcal{O}\left(\log\log n\right)$ multi-variable interventions are required on average to recover a causal graph over n variables once purely observational data is used to recover the "essential graph".

1.1.2 Effect estimation

Given a causal graph, estimating the strength of causality is similar to estimating the effect size in standard regression analysis [Gelman and Hill, 2006]. As we will see in Section 2.3, we can consider the problem of counterfactual reasoning in terms of two distinct conditional probability estimation tasks. It turns out these tasks are related (one factual and the second counterfactual) and tools from domain adaptation can be used for estimation [Johansson et al., 2016].

1.2 Observation vs Intervention

In contrast to standard statistical analysis based on observational data, causal analysis benefits from data that is collected after performing an intervention on the system of interest. One experimental design that allows us to study causal effects is the randomised control trial.

1.2.1 Learning from observational data

The goal of causal inference is to learn the effect of taking an action. We can do this directly via experimental approaches, however any given agent only has a limited capacity to manipulate the world. We are generating and storing data on almost every aspect of our lives at an unprecedented rate. As we incorporate sensors and robotics into our cities, homes, cars, everyday products and even our bodies, the breadth and scale of this data will only increase. However, only a tiny fraction of this data will be generated in a controlled way with the specific goal of answering a single question. An agent that can only learn from data when it had explicit control (or perfect knowledge of) the process by which that data was generated will be severely limited. This makes it critical that we develop effective methods that enable us to predict the outcome of an intervention in some system by observing, rather than acting on it. This is the problem of observational causal inference. The key feature that distinguishes observational from interventional data is that the learning agent does not control the action about which they are trying to learn.

1.2.2 Learning from interventions

The previous section focused on aspects of the problem of estimating the likely effect of an intervention from data gathered prior to making the intervention. There is an obvious alternative. Instead of trying to infer the outcome of an intervention from passive observations, one can intervene and see what happens. There are three key differences between observing a system and explicitly intervening in it. First, we determine the nature of the intervention and thereby control the data points used to estimate causal effects. Selecting data points optimally for learning is the focus of the optimal experimental design literature within statistics [Pukelsheim, 2006] and the active learning literature in machine learning [Settles, 2010]. Secondly, explicitly choosing interventions yields a perfect model of the probability with which each action is selected, given any context, allowing control over confounding bias. Finally, when we are intervening in a system we typically care about the impact of our actions on the system in addition to optimising learning. For example, in a drug trial, assigning people a sub-optimal treatment has real world costs. This leads to a trade-off between exploiting the best known
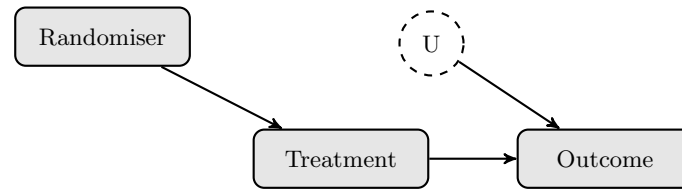
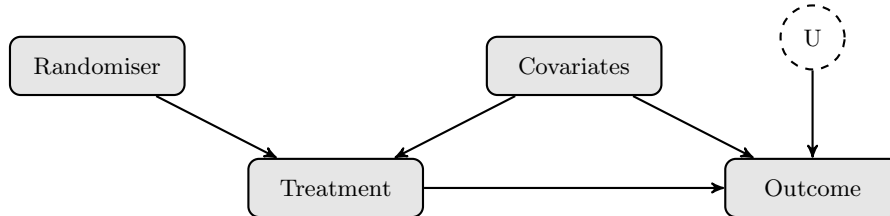Figure 1: causal network for a randomised experiment

Figure 2: causal network for a stratified randomised experiment if the probability an individual is assigned a given treatment depends on some covariates.

action so far and exploring alternative actions about which we are less certain. This exploration-exploitation trade-off lies at the heart of the field of reinforcement learning [Sutton and Barto, 1998].

1.3 Randomised experiments

Randomised controlled trials are often presented as the gold standard for determining causal effects. What is it about randomisation that makes it so important when it comes to causality? The graphical model for a randomised controlled experiment is shown in figure 1. If we assume perfect compliance (everyone takes the treatment that we select for them) then we have a perfect model for the treatment assignment process. Since treatment is assigned randomly, there can be no other variables that influence it and thus no confounding variables that affect both treatment and outcome.

Randomisation does not ensure target and control group are exactly alike. The more other features (observed or latent) influence the outcome, the more likely it is that there will be a significant difference in the joint distribution of these variables between the target and control groups in a finite data sample. However, the variance in the outcome, within both the target and control groups, also increases. The net result is increased variance (but not bias) in the estimate of causal effects.

Stratified randomised experiments address the issue of variance due to covariate imbalance by randomly allocating treatment conditional on covariates believed to influence the outcome of interest. If we stratify in such a way that the probability an instance receives a given treatment is independent of its covariates, for example, by grouping instances by each assignment to the covariates and then assigning treatment randomly with fixed probabilities, the causal graphical model in figure 1 still holds. We can then estimate the average causal effects directly from the differences in outcome across treatments. More complex stratification strategies can introduce a backdoor path from treatment to outcome via the covariates on which treatment is stratified, (figure 2). This necessitates that one condition on these covariates in computing the average causal effect in the same way as for estimating causal effects under ignorability. The key difference is that the propensity score is known, as it is designed by the experimenter, and there are guaranteed (rather than assumed) to be no latent confounding variables (that influence both treatment and outcome). See Imbens and Rubin [2015] for a discussion of the trade-offs between stratified versus completely random experiments.

The benefit provided by randomisation in breaking the link between the treatment variable and any latent confounders should not be understated. The possibility of unobserved confounders cannot be empirically ruled out from observational data [Pearl, 2000] (there is no test for confounding). This means causal estimates from non-experimental data are always subject to the criticism that an important confounder may have been overlooked or not properly adjusted for. However, randomised experiments do have some limitations.
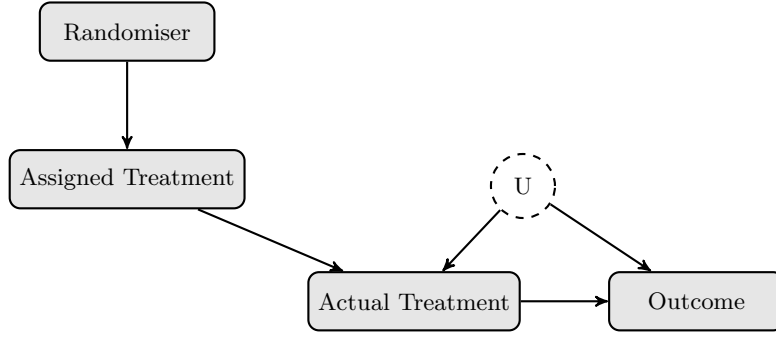
Figure 3: causal network for a randomised experiment with imperfect compliance

1.4 Limitations of randomised experiments

The idealised notion of an experiment represented by figure 1 does not capture the complexities of randomised experiments in practice. There may be imperfect compliance so that the treatment selected by the randomiser is not always followed, or output censoring in which the experimenter is not able to observe the outcome for all units (for example if people drop out). If compliance or attrition is not random, but associated with (potentially latent) variables that also affect the outcome, then the problem of confounding bias returns.[2] See figure 3 for a graphical model of a randomised experiment with imperfect compliance.

It is not always possible or ethical to conduct a randomised controlled trial, as is beautifully demonstrated by the paper of Smith and Pell [2003] on randomised cross-over trials of parachute use for the reduction of the mortality and morbidity associated with falls from large heights. When experimentation is possible, it is frequently difficult or expensive. This means experimental data sets are often much smaller than observational ones, limiting the complexity of models that can be explored. In addition, they are often conducted on a convenient, but unrepresentative, sample of the broader population of interest (for example first year university students). This can result in estimates with high *internal validity* [Campbell et al., 1963], in that they should replicate well in a similar population, but low *external validity* in that the results may not carry over to the general population of interest. The question of whether an experiment conducted on one population can be mapped to another is referred to as the transportability problem [Bareinboim and Pearl, 2013] and relies on very similar assumptions and arguments to causal inference and the do-calculus.

Finally, non-adaptive randomised experiments are not optimal from either an active or reinforcement learning perspective. In fact it is becoming increasingly common for adaptive clinical trials to be used as a study design [Bothwell et al., 2018]. As an experiment proceeds, information is obtained about the expectation and variance of each intervention (or treatment). Fixed experimental designs cannot make use of this information to select which intervention to try next. This results in both poorer estimates for a fixed number of experimental samples and more sub-optimal actions during the course of the experiment.

2 Four schools of causality

In this section, we consider limit the discussion to the case where we already know the causal graph and are only interested in estimating the causal effect from observational data. To simplify the presentation we only consider discrete random variables with finite number of states.

Observational causal inference aims to infer the outcome of an intervention in some system from data obtained by observing (but not intervening on) it. As previously mentioned, this is a form of transfer learning; we need to infer properties of the system post-intervention from observations of the

---

[2]Non-compliance is a problem if the goal is to estimate the causal effect of the treatment on the outcome but not if the goal is to estimate the causal effect of prescribing the treatment. The latter makes sense in a context where the process by which people decide whether to take the treatment they have been prescribed is likely to be the same if the treatment were made available more generally beyond the experimental trial.

system pre-intervention. Mapping properties from one system to another requires some assumptions about how these two systems are related, or in other words, a way of describing actions and how we anticipate a system will respond to them. Three key approaches have emerged: counterfactuals, structural equation models and causal Bayesian networks.

Counterfactuals [Rubin, 1974] were developed from the starting point of generalising from randomised trials to less controlled settings. They describe causal effects in terms of differences between counterfactual variables, what would happen if we took one action versus what would happen if we took another. Counterfactual assertions can be expressed very naturally in human languages and are prevalent in everyday conversations; "if I had worked harder I would have got better grades" and "she would have been much sicker if she hadn't taken antibiotics". Structural equation models have been developed and applied primarily within economics and related disciplines. They can be seen as an attempt to capture key aspects of the people's behaviour with mathematics. Questions around designing policies or interventions play a central role in economics. Thus they have transformed simultaneous equations into a powerful framework and associated set of methods for estimating causal effects. The is also a rich strand of work on using the assumptions that can be encoded in structural equation models, also known as functional causal models to discover the structure and direction of causal relationships - see for example Mooij et al. [2016]. Causal Bayesian networks [Pearl, 2000] are a more recent development and arise from the addition of a fundamental assumption about the meaning of a link to Bayesian networks. They inherit and leverage the way Bayesian networks encode conditional independencies between variables to localise the impact of an intervention in a system in a way that allows formalisation of the conditions under which causal effects can be inferred from observational data.

However, the literature on causal inference techniques remains split between the different frameworks. Much of the recent work on estimating causal effects within machine learning, as well as widely used methodologies such as propensity scoring, are described using the counterfactual framework. Methods developed within economics, in particular instrumental variable based approaches, or those requiring parametric or functional assumptions, are often based around structural equation models. This makes it worthwhile for researchers interested in causality to develop an understanding of all these viewpoints.

In the next sections, we describe causal Bayesian networks, counterfactuals and structural equation models: the problems they allow us to solve, the assumptions they rely on and how they differ. By describing all three frameworks, how they relate to one-another, and when they can be viewed as equivalent, we will make it easier for researchers familiar with one framework to understand the others and to transfer ideas and techniques between them. In order to demonstrate the notation and formalisms each framework provides, we will use them to describe the following simple examples.

Example 1. Suppose a pharmaceutical company wants to assess the effectiveness of a new drug on recovery from a given illness. This is typically tested by taking a large group of representative patients and randomly assigning half of them to a treatment group (who receive the drug) and the other half to a control group (who receive a placebo). The goal is to determine the clinical impacts of the drug by comparing the differences between the outcomes for the two groups (in this case, simplified to only two outcomes - recovery or non-recovery). We will use the variable X (1 = drug, 0 = placebo) to represent the treatment each person receives and Y (1 = recover, 0 = not recover) to describe the outcome.

Example 2. Suppose we want to estimate the impact on high school graduation rates of compulsory preschool for all four year olds. We have a large cross-sectional data set on a group of twenty year olds that records if they attended preschool, if they graduated high school and their parents socio-economic status (SES). We will let $X \in \{0,1\}$ indicate if an individual attended preschool, $Y \in \{0,1\}$ indicate if they graduated high school and $Z \in \{0,1\}$ represent if they are from a low or high SES background respectively.[3]

---

[3]There has been substantial empirical work on the effectiveness of early childhood education including a landmark randomised trial, the Perry Preschool project, which ran from 1962-1967 [Weikart and Others, 1970].

2.1 Causal Bayesian networks

Causal Bayesian networks are an extension of Bayesian networks. A Bayesian network is a graphical way of representing how a distribution factorises. Any joint probability distribution can be factorised into a product of conditional probabilities. There are multiple valid factorisations, corresponding to permutations of variable ordering.

$$P(X_1, X_2, X_3, ...) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)... \tag{1}$$

We can represent this graphically by drawing a network with a node for each variable and adding links from the variables on the right hand side to the variable on the left for each conditional probability distribution, see figure 4. If the factorisation simplifies due to conditional independencies between variables, this is reflected by missing edges in the corresponding network. There are multiple valid Bayesian network representations for any probability distribution over more than one variable, see figure 5 for an example.
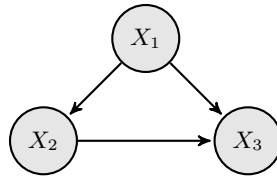

Figure 4: A general Bayesian network for the joint distribution over three variables. This network does not encode any conditional independencies between its variables and can thus represent any distribution over three variables.

The statement that a given graph G is a Bayesian network for a distribution P tells us that the distribution can be factorised over the nodes and edges in the graph. There can be no missing edges in G that do not correspond to conditional independencies in P, (the converse is not true: G can have extra edges). If we let $\mathbf{Pa}(X_i)$ represent the set of variables that are parents of the variable X_i in G then we can write the joint distribution as;

$$\mathrm{P}(X_1, ..., X_N) = \prod_{i=1...N} \mathrm{P}(X_i | \mathbf{Pa}(X_i)) \tag{2}$$

A causal Bayesian network is a Bayesian network in which a link $X_i \rightarrow X_j$, by definition, implies X_i causes X_j. This means an intervention to change the value of X_i can be expected to affect X_j, but interventions on X_j will not affect X_i. We need some notation to describe interventions and represent distributions over variables in the network after an intervention. We use the do operator introduced by Pearl [Pearl, 2000].

Definition 1. The do-notation

- $do(X = x)$ denotes an intervention that sets the random variable(s) X to x.

- $\mathrm{P}(Y|do(X))$ is the distribution of Y conditional on an *intervention* that sets X. This notation is somewhat overloaded. It may be used to represent a probability distribution/mass function or a family of distribution functions depending on whether the variables are discrete or continuous and whether or not we are treating them as fixed. For example, it could represent

    - the probability $\mathrm{P}(Y = 1|do(X = x))$ as a function of x,
    - the probability mass function for a discrete Y : $\mathrm{P}(Y|do(X = x))$,
    - the probability density function for a continuous Y : $f_Y(y|do(X = x))$,
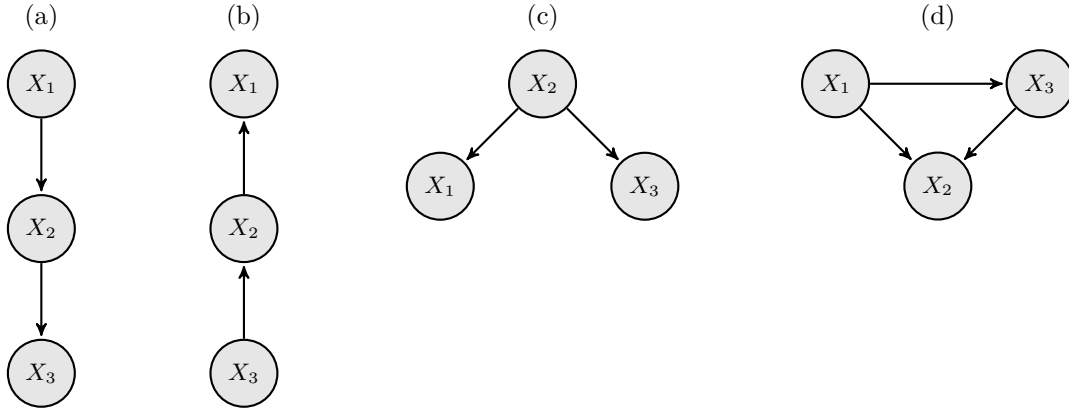   - a family of density/mass function for Y parameterised by x.

Figure 5: Some valid Bayesian networks for a distribution P over (X_1, X_2, X_3) in which X_3 is conditionally independent of X_1 given X_2, denoted $X_3 \perp\!\!\!\perp X_1|X_2$. Graphs (a), (b) and (c) are all a *perfect map* for P as the graphical structure implies exactly the same set of independencies exhibited by the distribution. Graph (d), like figure 4 does not imply any conditional independencies, and is thus a valid (but not very useful) Bayesian network representation for any distribution over three variables.

Where the distinction is important and not clear from context we will use one of the more specific forms above.

Theorem 2 (Truncated product formula [Pearl, 2000]). *If G is a causal network for a distribution P defined over variables $X_1...X_N$, then we can calculate the distribution after an intervention where we set $Z \subset X$ to z, denoted $do(Z = z)$ by dropping the terms for each of the variables in Z from the factorisation given by the network. Let $\mathbf{Pa}(X_i)$ denote the parents of the variable X_i in G.*

$$\mathrm{P}\left(X_1...X_N|do(Z = z)\right) = \mathbb{1}\{Z = z\} \prod_{X_i \notin Z} \mathrm{P}\left(X_i|\mathbf{Pa}(X_i)\right) \tag{3}$$

Theorem 2 does not hold for standard Bayesian networks because there are multiple valid networks for the same distribution. The truncated product formula will give different results depending on the selected network. The result is possible with causal Bayesian networks because it follows directly from the assumption that the direction of the link indicates causality. In fact, from the interventionist viewpoint of causality, the truncated product formula defines what it means for a link to be causal.

Returning to example 1, and phrasing our query in terms of interventions; what would the distribution of outcomes look like if everyone was treated $\mathrm{P}\left(Y|do(X = 1)\right)$, relative to if no one was treated $\mathrm{P}\left(Y|do(X = 0)\right)$? The treatment X is a potential cause of Y, along with other unobserved variables, such as the age, gender and the disease subtype of the patient. Since X is assigned via deliberate randomisation, it cannot be affected by any latent variables. The causal Bayesian network for this scenario is shown in figure 6. This network represents the (causal) factorisation $\mathrm{P}\left(X, Y\right) = \mathrm{P}\left(X\right)\mathrm{P}\left(Y|X\right)$, so from equation (3), $\mathrm{P}\left(Y|do(X)\right) = \mathrm{P}\left(Y|X\right)$. In this example, the interventional distribution is equivalent to the observational one.
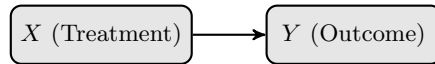

Figure 6: Causal Bayesian network for example 1

In example 2 we are interested in $\mathrm{P}\left(Y|do(X = 1)\right)$, the expected high-school graduation rate if we introduce universal preschool. We could compare it to outlawing preschool $\mathrm{P}\left(Y|do(X = 0)\right)$ or the current status quo $\mathrm{P}\left(Y\right)$. It seems reasonable to assume that preschool attendance affects the likelihood of high school graduation [4] and that parental socio-economic status would affect *both* the

---

[4]The effect does not have to be homogeneous, it may depend non-linearly on characteristics of the child, family and school.
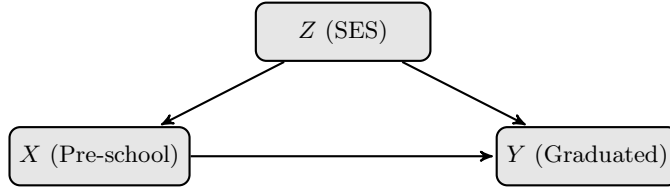
Figure 7: Causal Bayesian network for example 2

likelihood of preschool attendance and high school graduation. If we assume that socio-economic status is the only such variable (nothing else affects both attendance *and* graduation), we can represent this problem with the causal Bayesian network in figure 7. In this case, the interventional distribution is not equivalent to the observational one. If parents with high socio-economic status are more likely to send their children to preschool and these children are more likely to graduate high school regardless, comparing the graduation rates of those who attended preschool with those who did not will overstate the benefit of preschool. To obtain the interventional distribution we have to estimate the impact of preschool on high school graduation for each socio-economic level separately and then weight the results by the proportion of the population in that group,

$$\mathrm{P}\left(Y|do(X=1)\right) = \sum_{z \in Z} \mathrm{P}\left(Y|X=1,Z\right)\mathrm{P}\left(Z\right) \tag{4}$$

We have seen from these two examples that the expression to estimate the causal effect of an intervention depends on the structure of the causal graph. There is a very powerful and general set of rules that specifies how we can transform observational distributions into interventional ones for a given graph structure. These rules are referred to as the Do-calculus [Pearl, 2000].

A causal Bayesian network represents much more information than a Bayesian network with identical structure. A causal network encodes all possible interventions that could be specified with the do-notation. For example, if the network in figure 7 were an ordinary Bayesian network and all the variables were binary, the associated distribution could be described by seven parameters. The equivalent causal Bayesian network additionally represents the post-interventional distributions for six possible single variable interventions and twelve possible two variable interventions. Encoding all this information without the assumptions implicit in the causal Bayesian network would require an additional thirty parameters.[5]

Causal Bayesian networks are Bayesian networks, so results that apply to Bayesian networks carry directly across: the local Markov property states that variables are independent of their non-effects given their direct causes. The global Markov property and d-separation also hold in causal networks. D-separation, which characterises which conditional independencies must hold in any distribution that can be represented by a given Bayesian network G, is key to many important results and algorithms for causal inference.

2.2   Limitations of causal Bayesian networks

A number of criticisms have been levelled at this approach to modelling causality. One is that the definition of an intervention only in terms of setting the value of one or more variables is too precise and that any real world intervention will affect many variables in complex and non-deterministic ways [Rickles, 2009, Cartwright, 2007]. However, by augmenting the causal graph with additional variables that model how interventions may take effect, the deterministic do operator can model more complex interventions. For example, in the drug treatment case, we assumed that all subjects complied, taking the treatment or placebo as assigned by the experimenter. But, what if some people failed to take the prescribed treatment? We can model this within the framework of deterministic interventions

---

[5]After each single variable intervention we have a distribution over two variables, which can be represented by three parameters. After each two variable intervention, we have a distribution over one variable which requires one parameter. This takes us to a total of $6 \times 3 + 12 \times 1 = 30$ additional parameters.

by adding a node representing what they were prescribed (the intervention) which probabilistically influences the treatment they actually receive (figure 8). Note that the fact that we no longer directly assign the treatment opens the possibility that an unobserved latent variable could affect both the actual treatment taken and the outcome.
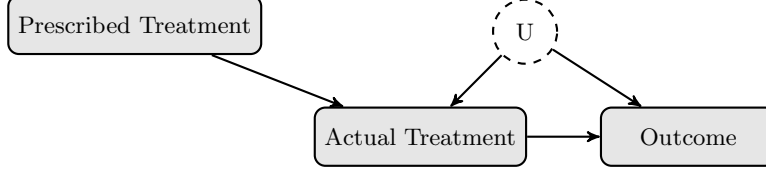

Figure 8: Randomised treatment with imperfect compliance

Another key issue with causal Bayesian networks is that they cannot handle cyclic dependencies between variables. Such feedback loops are common in real-life systems, for example the relationship between supply and demand in economics or predator and prey in ecology. We might regard the underlying causal mechanisms in these examples to be acyclic; the number of predators at one time influences the number of prey in the next period and so on. However, if our measurements of these variables must be aggregated over time periods that are longer than the scale at which these interactions occur, the result is a cyclical dependency. Even were we able to measure on shorter timescales, there might then not be sufficient data on each variable for inference. Such problems have mostly been studied within the dynamical systems literature, typically focusing on understanding the stationary or equilibrium state of the system and making very specific assumptions about functional form in order to make problems tractable. Poole and Crowley [2013] compare the equilibrium approach to reasoning about cyclic problems with structural equation models, which we discuss in section 2.4 and that can be seen as Bayesian causal networks with additional functional assumptions.

2.3 Counterfactuals

The Neyman-Rubin model [Rubin, 1974, 1978, Rosenbaum and Rubin, 1983, Rubin, 2005, 2008] defines causality in terms of potential outcomes, or counterfactuals. Counterfactuals are statements about imagined or alternate realities, are prevalent in everyday language and may play a role in the development of causal reasoning in humans [Weisberg and Gopnik, 2013]. Causal effects are differences in counterfactual variables: what the difference is between what would have happened if we did one thing versus what would have happened if we did something else.

In example 1, the causal effect of the drug relative to placebo for person i is the difference between what would have happened if they were given the drug, denoted y_i^1 versus what would have happened if they got the placebo, y_i^0. The fundamental problem of causal inference is that we can only observe one of these two outcomes, since a given person can only be treated or not treated. The problem can be resolved if, instead of people, there are units that can be assumed to be identical or that will revert exactly to their initial state some time after treatment. This type of assumption often holds to a good approximation in the natural sciences and explains why researchers in these fields are less concerned with causal theory.

Putting aside any estimates of individual causal effects, it is possible to learn something about the distributions under treatment or placebo. Let Y^1 be a random variable representing the potential outcome if treated. The distribution of Y^1 is the distribution of Y if everyone was treated. Similarly Y^0 represents the potential outcome for the placebo. The difference between the probability of recovery, across the population, if everyone was treated and the probability of recovery if everyone received the placebo is $P(Y^1) - P(Y^0)$. We can estimate (from an experimental or observational study):

- $P(Y = 1 | X = 1)$, the probability that those who took the treatment will recover

- $P(Y = 1 | X = 0)$, the probability that those who were *not* treated will recover

Now, for those who took the treatment, the outcome *had* they taken the treatment Y^1 is the same as the observed outcome. For those who did not take the treatment, the observed outcome is the same as the outcome *had* they not taken the treatment. Equivalently stated:

$$P\left(Y^0|X=0\right) = P\left(Y|X=0\right)$$
$$P\left(Y^1|X=1\right) = P\left(Y|X=1\right)$$

If we assume $X \perp\!\!\!\perp Y^0$ and $X \perp\!\!\!\perp Y^1$:

$$P\left(Y^1\right) = P\left(Y^1|X=1\right) = P\left(Y|X=1\right)$$
$$P\left(Y^0\right) = P\left(Y^0|X=0\right) = P\left(Y|X=0\right)$$

This implies the counterfactual distributions are equivalent to the corresponding conditional distributions and, for a binary outcome Y, the causal effect is,

$$P\left(Y^1\right) - P\left(Y^0\right) = P\left(Y|X=1\right) - P\left(Y|X=0\right)$$

The assumptions $X \perp\!\!\!\perp Y^1$ and $X \perp\!\!\!\perp Y^0$ are referred to as ignorability assumptions [Rosenbaum and Rubin, 1983]. They state that the treatment each person receives is independent of whether they would recover if treated and if they would recover if not treated. This is justified in example 1 due to the randomisation of treatment assignment. In general the treatment assignment will not be independent of the potential outcomes. In example 2, the children from wealthy families could be more likely to attend preschool but also more likely to do better in school regardless, i.e $X \not\perp\!\!\!\perp Y^0$ and $X \not\perp\!\!\!\perp Y^1$. A more general form of the ignorability assumption is to identify a set of variables Z such that $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$.

Theorem 3 (Ignorability [Rosenbaum and Rubin, 1983, Pearl, 2000]). *If $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$,*

$$P\left(Y^1\right) = \sum_{z \in Z} P\left(Y|X=1,Z\right)P\left(Z\right) \tag{5}$$

$$P\left(Y^0\right) = \sum_{z \in Z} P\left(Y|X=0,Z\right)P\left(Z\right) \tag{6}$$

Assuming that within each socio-economic status level, attendance at preschool is independent of the likelihood of graduating high-school had a person attended, then the average rate of high-school graduation given a universal preschool program can be computed from equation 5. Note, that this agrees with the weighted adjustment formula in equation 4.

Another assumption introduced within the Neyman-Rubin causal framework is the Stable Unit Treatment Value Assumption (SUTVA) [Rubin, 1978]. This is the assumption that the potential outcome for one individual (or unit) does not depend on the treatment assigned to another individual. As an example of a SUTVA violation, suppose disadvantaged four year olds were randomly assigned to attend preschool. The subsequent school results of children in the control group, who did not attend, could be boosted by the improved behaviour of those who did and who now share the classroom with them. SUTVA violations would manifest as a form of model misspecification in causal Bayesian networks.

There are objections to counterfactuals arising from the way they describe alternate universes that were never realised. In particular, statements involving joint distributions over counterfactual variables may not be able to be validated empirically Dawid [2000]. One way of looking at counterfactuals is as a natural language short hand for describing highly specific interventions like those denoted by the do-notation. Rather than talking about the distribution of Y given we intervene to set $X = x$ and hold everything else about the system constant we just say what would the distribution of Y be had X been x. This is certainly convenient, if rather imprecise. However, the ease with which we can make statements with counterfactuals that cannot be tested with empirical data warrants careful attention. It is important to be clear what assumptions are being made and whether or not they could be validated (at least in theory).

2.4 Structural Equation models

Structural equation models (SEMs) describe a deterministic world, where some underlying mechanism or function determines the output of any process for a given input. The mechanism (but not the output) is assumed to be independent of what is fed into it. Uncertainties are not inherent but arise from unmeasured variables. Linear structural equation models have a long history for causal estimation Wright [1921], Haavelmo [1943]. More recently, they have been formalised, generalised to the non-linear setting and connected to developments in graphical models to provide a powerful causal framework [Pearl, 2000].

Mathematically, each variable is a deterministic function of its direct causes and a noise term that captures unmeasured variables. The noise terms are required to be mutually independent. If there is the possibility that an unmeasured variable influences more than one variable of interest in a study, it must be modelled explicitly as a latent variable. Structural equation models can be represented visually as a network. Each variable is a node and arrows are drawn from causes to their effects. Figure 9 illustrates the SEM for example 1.
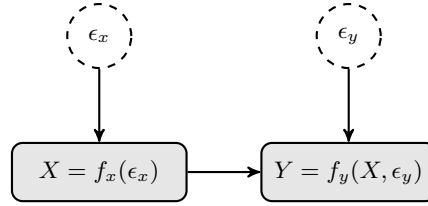

Figure 9: SEM for example 1

This model encodes the assumption that the outcome y_i for an individual i is caused solely by the treatment x_i they receive and other factors ϵ_{y_i} that are independent of X. This is justifiable on the grounds that X is random. The outcome of a coin flip for each patient should not be related to any of their characteristics (hidden or otherwise). Note that the causal graph in figure 9 is identical to that of the Bayesian network for the same problem (figure 6). The latent variables ϵ_x and ϵ_y are not explicitly drawn in figure 6 as they are captured by the probabilistic nature of the nodes in a Bayesian network.

Taking the *action* $X = 1$ corresponds to replacing the equation $X = f_x(\epsilon_x)$ with $X = 1$. The function f_y and distribution over ϵ_y does not change. This results in the interventional distribution, [6]

$$\mathrm{P}\left(Y = y | do(X = 1)\right) = \sum_{\epsilon_y} \mathrm{P}\left(\epsilon_y\right) \mathbb{1}\{f_y(1, \epsilon_y) = y\} \tag{7}$$

The observational distribution of Y given X is,

$$\mathrm{P}\left(Y = y | X = 1\right) = \sum_{\epsilon_x}\sum_{\epsilon_y} \mathrm{P}\left(\epsilon_x | X = 1\right) \mathrm{P}\left(\epsilon_y | \epsilon_x\right) \mathbb{1}\{f_y(1, \epsilon_y) = y\} \tag{8}$$

$$= \sum_{\epsilon_y} \mathrm{P}\left(\epsilon_y\right) \mathbb{1}\{f_y(1, \epsilon_y) = y\}, \text{ as } \epsilon_x \perp\!\!\!\perp \epsilon_y \tag{9}$$

The interventional distribution is the same as the observational one. The same argument applies to the intervention $do(X = 0)$, and so the causal effect is simply the difference in observed outcomes as found via the causal Bayesian network and counterfactual approaches.

The SEM for example 2 is shown in figure 10. Intervening to send all children to preschool replaces the equation $X = f_x(Z, \epsilon_x)$ with $X = 1$, leaving all the other functions and distributions in the model unchanged.

---

[6]We have assumed the variables are discrete only for notational convenience

$$P\left(Y = y|do(X = 1)\right) = \sum_z \sum_{\epsilon_y} P\left(z\right) P\left(\epsilon_y\right) \mathbb{1}\{f_y(1, z, \epsilon_y) = y\} \tag{10}$$

$$= \sum_z P\left(z\right) \underbrace{\sum_{\epsilon_y} P\left(\epsilon_y\right) \mathbb{1}\{f_y(1, z, \epsilon_y) = y\}}_{P(Y=y|X=1, Z=z)} \tag{11}$$

Equation 11 corresponds to equations 4 and 5. It is not equivalent to the observational distribution given by:

$$P\left(Y = y|X = 1\right) = \sum_z \sum_{\epsilon_y} P\left(z|X = 1\right) P\left(\epsilon_y\right) \mathbb{1}\{f_y(1, z, \epsilon_y) = y\} \tag{12}$$
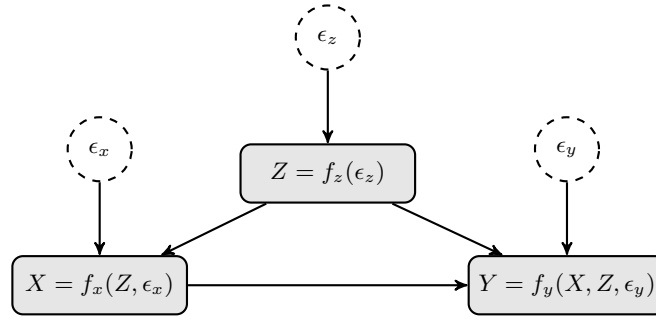

Figure 10: SEM for example 2

Structural equation models are generally applied with strong constraints on the functional form of the relationship between the variables and noise, which is typically assumed to be additive, $X_i = f_i(\cdot) + \epsilon_i$. A structural equation model with N variables resembles a set of N simultaneous equations, with each variable playing the role of the dependent (left hand side) variable in one equation. However a SEM is, by definition, more than a set of simultaneous equations. By declaring it to be structural, we are saying that it represents *causal* assumptions about the relationships between variables. When visualised as a network, the absence of an arrow between two variables encodes the assumption that one does not cause the other. The similarity between the notation used to describe and analyse structural equation models and simultaneous equations, combined with a reluctance to make explicit statements about causality, has led to some confusion in the interpretation of SEMs [Heckman and Pinto, 2015, Pearl, 2000].

2.5 Comparing and unifying the models

Remarkably for models developed relatively independently in fields with very different approaches and problems, causal Bayesian networks, counterfactuals and structural equation models can be nicely unified for interventional queries (those that can be expressed with the do-notation) [Pearl, 2000]. These queries, and the assumptions required to answer them, can be mapped between the frameworks in a straightforward way, allowing techniques developed within one framework to be immediately applied within another. If the network for a structural equation model is acyclic, that is if starting from any node and following edges in the direction of the arrows you cannot return to the starting point, then it implies a recursive factorisation of the joint distribution over its variables. In other words, the network is a causal Bayesian network. All of the results that apply to causal Bayesian networks also apply to acyclic structural equation models. Taking an action that sets a variable to a specific value equates to replacing the equation for that variable with a constant. This corresponds to dropping a term in the factorisation and the truncated product formula (equation 3). Thus, the interventional query $P(Y|do(X))$ is identical in these two frameworks. We can also connect this to counterfactuals via:

$$P\left(Y^0\right) \equiv P(Y|do(X = 0))$$
$$P\left(Y^1\right) \equiv P(Y|do(X = 1))$$
(13)

The assumption $\epsilon_X \perp\!\!\!\perp \epsilon_Y$, stated for our structural equation model, translates to $X \perp\!\!\!\perp (Y^0, Y^1)$ in the language of counterfactuals. When discussing the counterfactual model, we made the slightly weaker assumption:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1$$
(14)

It is possible to relax the independence of errors assumption for SEMs to correspond exactly with the form of equation (14) without losing any of the power provided by d-separation and graphical identification rules [Richardson and Robins, 2013]. The correspondence between the models for interventional queries (those that can be phrased using the do-notation) makes it straightforward to combine key results and algorithms developed within any of these frameworks. For example, you can draw a causal graphical network to determine if a problem is identifiable and which variables should be adjusted for to obtain an unbiased causal estimate. Then use propensity scores [Rosenbaum and Rubin, 1983] to estimate the effect. If non-parametric assumptions are insufficient for identification or lead to overly large uncertainties, you can specify additional assumptions by phrasing your model in terms of structural equations. The frameworks do differ when it comes to causal queries that involve joint or nested counterfactuals and cannot be expressed with the do-notation. These types of queries arise in the study of mediation [Pearl, 2014, Imai et al., 2010, VanderWeele and Hernández-Diaz, 2011] and in legal decisions, particularly on issues such as discrimination [Pearl, 2000]. The graphical approach to representing causal knowledge can be extended to cover these types of questions via Single World Intervention Graphs [Richardson and Robins, 2013], which explicitly represent counterfactual variables in the graph.

In practice, differences in focus and approach between the fields in which each model dominates eclipse the actual differences in the frameworks. The work on causal graphical models [Pearl, 2000, Spirtes et al., 2000] focuses on asymptotic, non-parametric estimation and rigorous theoretical foundations. The Neyman-Rubin framework builds on the understanding of randomised experiments and generalises to quasi-experimental and observational settings, with a particular focus on non-random assignment to treatment. Treatment variables are typically discrete (often binary). This research emphasises estimation of average causal effects and provides practical methods for estimation, in particular, propensity scores; a method to control for multiple variables in high dimensional settings with finite data [Rosenbaum and Rubin, 1983]. In economics, inferring causal effects from non-experimental data to support policy decisions is central to the field. Economists are often interested in more informative measures of the distribution of causal effects than the mean and make extensive use of structural equation models, generally with strong parametric assumptions [Heckman, 2008]. The central approach to estimation is regression - which naturally handles continuous variables while discrete variables are typically encoded as indicator variables. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models. However, these differences are not fundamental to the frameworks. Functional assumptions can be specified on the conditional distributions of (causal) Bayesian networks, counterfactuals can readily represent continuous treatments (eg Y^x), and structural equation models can represent complex non-linear relationships between both continuous and discrete variables.

2.6 Granger causality

A discussion of approaches to (observational) causal inference would not be complete without a mention of Granger causality, [Granger, 1969]. The fundamental idea behind Granger causality is to leverage the assumption that the future does not cause the past to test the existence and direction of a causal link between two time series. The basic approach is to test, for a pair of time series variables X and Y, if $Y_t \perp\!\!\!\perp (X_1, ..., X_{t-1})|(Y_1, ..., Y_{t-1})$ - that is if the history of X helps to predict Y given the history of Y. The original formulation considered only pairs of variables and linear causal relationships but recent work has generalised the key idea to multiple variables and non-linear relationships. Unlike the

previous models we have discussed, Granger causality does not provide us with a means to specify our assumptions about the causal structure between variables. Rather it aims to infer the causal structure of a structural equation model from observational data - subject to some assumptions. In some sense, Granger causality is closer to standard prediction tasks in machine learning than causal analysis.

References

Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, 2002.

Elias Bareinboim and Judea Pearl. Transportability from Multiple Environments with Limited Experiments. *Advances in Neural Information Processing Systems*, pages 1–9, 2013. URL http://papers.nips.cc/paper/5062-transportability-from-multiple-environments-with-limited-experiment

Laura E Bothwell, Jerry Avorn, Nazleen F Khan, and Aaron S Kesselheim. Adaptive design clinical trials: a review of the literature and clinicaltrials.gov. *BMJ Open*, 8(2), 2018. ISSN 2044-6055. doi: 10.1136/bmjopen-2017-018320. URL http://bmjopen.bmj.com/content/8/2/e018320.

DT Campbell, JC Stanley, and NL Gage. *Experimental and quasi-experimental designs for research*. Houghton Mifflin, Boston, 1963. URL http://moodle.technion.ac.il/pluginfile.php/219643/mod{_}resource/content/0/Campbell{_}and{_}Stanley{

Nancy Cartwright. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press, 2007.

AP Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 2000. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210.

Frederick Eberhardt. Causal Discovery as a Game. In *NIPS Causality: Objectives and Assessment*, pages 87–96, 2010.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *UAI*, 2005. ISBN 0-9749039-1-4. URL http://arxiv.org/abs/1207.1389.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Andrew Gelman and Guido Imbens. Why ask why? forward causal inference and reverse causal questions. Technical Report 19614, The National Bureau of Economic Research, 2013.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424, 1969. ISSN 00129682. doi: 10.2307/1912791. URL http://www.jstor.org/stable/1912791?origin=crossref.

T Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12, 1943. URL http://www.jstor.org/stable/1905714.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, may 2014. doi: 10.1016/j.ijar.2013.11.007. URL http://arxiv.org/abs/1205.4174http://dx.doi.org/10.1016/j.ijar.2013.11.007.

James Heckman and Rodrigo Pinto. Causal analysis after Haavelmo. *Econometric Theory*, 31(01):115–151, 2015.

James JJ Heckman. Econometric causality. *International Statistical Review*, 2008. doi: 10.1073/pnas.0703993104. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2007.00024.x/full.

Huining Hu, Zhentao Li, and Adrian R Vetta. Randomized Experimental Design for Causal Graph Discovery. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2339–2347. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5535-randomized-experimental-design-for-causal-graph-discovery.pdf.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51–71, feb 2010. ISSN 0883-4237. doi: 10.1214/10-STS321. URL http://projecteuclid.org/euclid.ss/1280841733.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

Finnian Lattimore. *Learning how to act: making good decisions with machine learning*. PhD thesis, Australian National University, 2018.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL http://jmlr.org/papers/v17/14-518.html.

Judea Pearl. *Causality: models, reasoning and inference*. MIT Press, Cambridge, 2000.

Judea Pearl. Interpretation and Identification of Causal Mediation. *Psychological methods*, jun 2014. ISSN 1939-1463. doi: 10.1037/a0036434. URL http://www.ncbi.nlm.nih.gov/pubmed/24885338.

Judea Pearl and Dana Mackenzie. *The book of why: The new science of cause and effect*. Basic Books, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

David Poole and Mark Crowley. Cyclic causal models with discrete variables: Markov chain equilibrium semantics and sample ordering. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1060–1068, 2013. ISSN 10450823.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Thomas S Richardson and James M Robins. Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(128), 2013. URL https://iriss.stanford.edu/sites/all/files/dsi/jamesrobinspaper.pdf.

Dean Rickles. Causality in complex interventions. *Medicine, Health Care and Philosophy*, 12(1):77–90, 2009.

PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. URL http://biomet.oxfordjournals.org/content/70/1/41.short.

DB Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 1974. URL http://psycnet.apa.org/journals/edu/66/5/688/.

DB Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 1978. URL http://www.jstor.org/stable/2958688.

DB Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, mar 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL http://www.tandfonline.com/doi/abs/10.1198/016214504000001880.

DB Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, sep 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS187. URL http://projecteuclid.org/euclid.aoas/1223908042.

Burr Settles. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin-Madison, 2010.

Gordon C S Smith and Jill P Pell. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ: British Medical Journal*, 327(7429):1459, 2003.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x. URL http://www.applied-informatics-j.com/content/3/1/3.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Tyler J VanderWeele and Sonia Hernández-Diaz. Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatric and perinatal epidemiology*, 25(2):111–5, mar 2011. ISSN 1365-3016. doi: 10.1111/j.1365-3016.2010.01175.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21281323.

David P Weikart and Others. Longitudinal Results of the Ypsilanti Perry Preschool Project. Final Report. Volume II of 2 Volumes., 1970.

Deena S Weisberg and Alison Gopnik. Pretense, counterfactuals, and Bayesian causal models: why what is not real really matters. *Cognitive science*, 37(7):1368–81, 2013. ISSN 1551-6709. doi: 10.1111/cogs.12069. URL http://www.ncbi.nlm.nih.gov/pubmed/23915198.

S Wright. Correlation and causation. *Journal of agricultural research*, 1921. URL http://www.ssc.wisc.edu/soc/class/soc952/Wright/Wright{_}CorrelationandCausation.pdf.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-Scale Kernel Methods for Independence Testing. *Statistics and Computing*, 28(1):113–130, January 2018.