10-708: Probabilistic Graphical Models 10-708, Spring 2018

4: Causal Graphic Models

Lecturer: Kun Zhang *Scribes: Momin M. Malik and Naji Shajarisales*

1 Introduction

How do we use causal graphs and causal representations? How do we discover causal information from data? How do we do machine learning from the causal perspective? What's the difference between conditional independence and causality? We know that formally X and Y are **associated** iff $\exists\ x_1 \neq x_2$ such that $P(Y|X = x_1) \neq P(Y|X = x_2)$. This does necessary mean a causal relationship; causation is stronger than correlation/dependence.

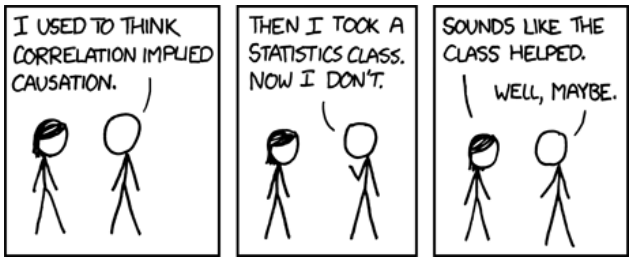

Figure 1: The obligatory XKCD for causation. `https://xkcd.com/552/`.

The classic statement (fig. 1) seems to be **too** strong; today we know that in many cases we can actually use observational data to recover causal relations. How can we make sure a dependence is causal and not just a mere dependence? This question has been pondered for a long time—for at least 2500 years, since Aristotle.

The formal definition of causation can be done in terms of manipulation: X is a **cause** of Y iff $\exists\ x_1 \neq x_2$ such that $P(Y|\text{do}(X = x_1)) \neq P(Y|\text{do}(X = x_2))$, where here the "do" operator is a formal manipulation to be defined later, but intuitively it means *setting* the value of a random variable to a specific value.

1.1 Spurious Causal Conclusions

Example 1: An article looking at grunting in tennis. They found a 4.9% enhancement in serve velocity among players who grunted [5], but looking more closely one can see paper does not offer any evidence for causality. Of course one can think of possibility of reverse causation, or common cause (which in this case would be from *trying hard*).

Example 2: In a very good study [6], large-scale psychological differences within China can be **explained by** rice versus wheat agriculture. People who farm rice and farm wheat have psychological differences. Clearly, we cannot do any experiment: we have to analyze observations. They collected data points all over the world, or all over China. There is no other way but analyzing data: from that, they claim a very strong conclusion. We will come back to this and see that while strong, this claim is reasonable.

Example 3: There is a high correlation between annual chocolate consumption of countries and the number of their Nobel laureate (fig. 2) [3]. You really want to see why this is the case. A bizarre causal conclusion would lead to distributing chocolate everywhere to increase Nobel laureates in one's country. But this is indeed bizarre. We really want to find the causal relationship to avoid such conclusions.

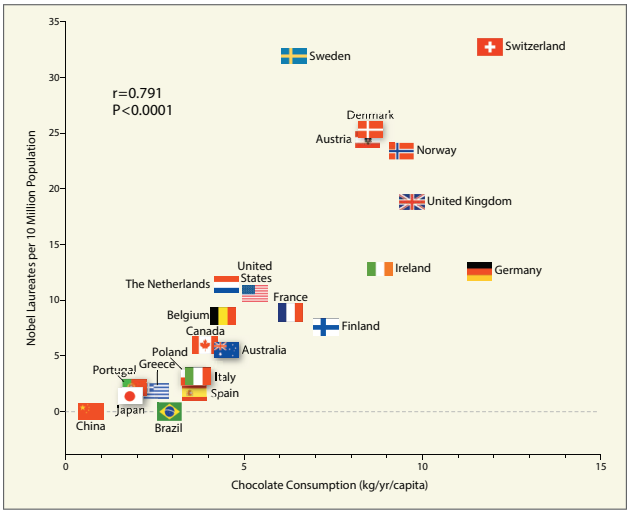

Figure 2: Correlation between countries annual per capita chocolate consumption and the number of Nobel Laureates per 10 million population, via [3].

2 Outline

1. How can we benefit from causal thinking? What are the underlying process behind data/observations?
2. Causal representations. We use graphical models to represent causality, but causal representation is stronger than just a Bayesian network.
3. Interventions. We can plan interventions, and show how we will make use of interventions. Here, there are two main tasks:
   (a) Identification of causal effects. Are causal effects identifiable? How?
   (b) Causal discovery. We have been working on this probelm for very long. It's a very interesting problem; people are making quick progress, and it's very exciting. We will talk about this later, as well as how to make use of causality for machine learning.

   Also, we will briefly discuss how to understand/model causal *cycles* which are ubiquitous in real-world problems. We always have suchh cycles, which is why causality is so complicated.

3 Causal Thinking

3.1 Forms of dependence

We will review three causal structures that produce dependencies between variables.

3.1.1 Common cause

This is an example from R. A. Fisher in 1950s:[1] there is a high correlation between the color of fingers and lung cancer. But if you want to change the incidence of lung cancer, you can't rely on this dependence: smoking might be main cause of both yellow fingers and lung cancer (fig. 3).

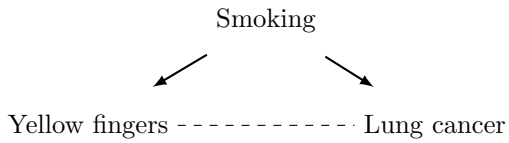

Figure 3: A causal DAG for relationship between whether somebody has yellow fingers, whether that person has lung cancer and whether they smoke.

When one is after *controlling* a phenomenon, such as advertising for a product, suggesting a treatment, or designing a policy, one wants to make sure they understand causal representations and causal process. For example, if you advertise a new product, you want to know if and how that advertisement will change people's minds.

3.1.2 Causal relations

The following table shows the outcome of applying two different kind of treatments to cure kidney stones, from a real (observational) study [1] used as an example of Simpson's paradox [2].

	Treatment A	Treatment B
Small Stones	(Group 1) **93% (81/87)**	(Group 2) 87% (234/270)
Large Stones	(Group 3) **73% (192/263)**	(Group 4) 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Another way of representing this is in a diagram (fig. 4).

Based on these observations, the 'paradoxical' question, is which method should we choose if a patient shows up with a small kidney stone? Combining patients together, one can see that treatment B is more effective overall. But regardless of whether stone is small or large, we prefer A. This is known as **Simpson's paradox**: all subgroups exhibit the opposite trend from that exhibited in the total population.

Another example of Simpson's paradox is the relation between exercise and cholesterol. The amount of exercise one does and the cholesterol level of their body are positively correlated. When the age is not considered, the most exercise one does seems to say the more cholesterol one has. But dividing the result into different age groups as is shown in figure 5 left we can see indeed the correlation is negative. How can we explain this?

Getting back to the kidney stone treatment problem, the key observation is that how you divide patients

[1] Scribe's note: Despite his many contributions to statistics, Fisher was an avid pipe smoker who, until the end of his life, resisted the causal relationship between smoking and cancer. He happily took money from the tobacco industry to produce materials arguing against any connection.

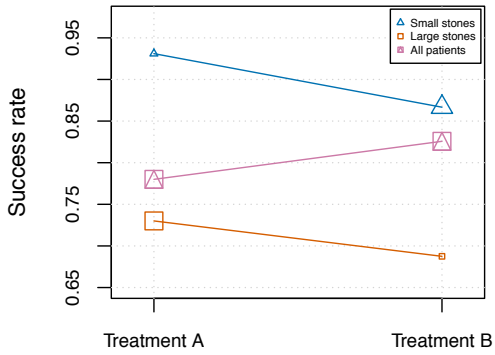

Figure 4: Comparison of total observational outcomes to outcomes conditioned on stone size.

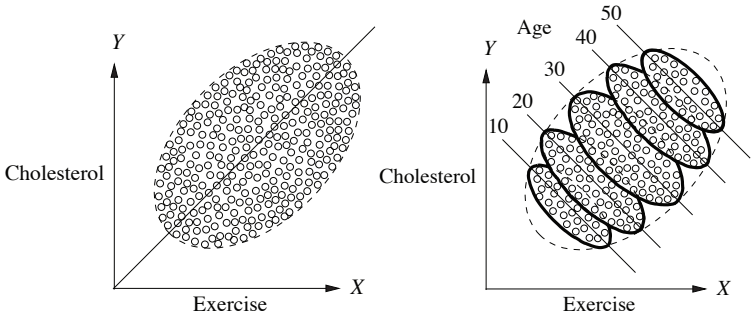

Figure 5: Plot of cholesterol in blood w.r.t. to amount of exercise (left), and the same relationship clustered by age group (right). Images are from [4].

into groups influences treatment and recovery. There is a missing common cause; and with such a common cause, the dependence pattern can be almost arbitrary (fig. 6).

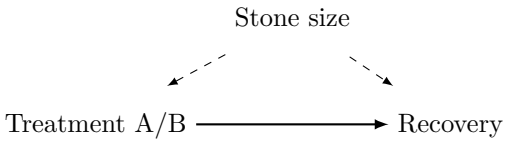

Figure 6: The DAG explaining the causal relation between the variables of kidney stone problem.

Thus, if you were making recommendations about a treatment, using correlation alone would lead to worse outcomes.

3.1.3 V-structures

50 years ago, female college students were smarter than male ones on average. Why? A reasonable *causal* explanation is that only the smartest women were admitted to the college, and data for this assessment was gathered among college students (fig. 7).

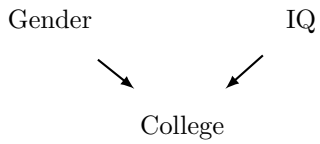

Figure 7: A "v-structure" for college students in the 1950s.

Another example is the "Monty Hall" problem[2]: the money and your original choice are independent. But, which door is opened by Monty is a common effect.

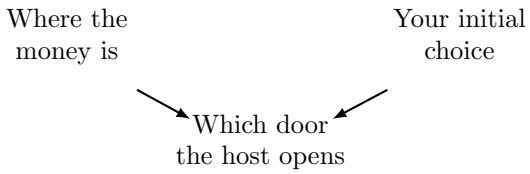

Figure 8: A "v-structure" in the Monty Hall problem.

V-structures are very important, since whenever we deal with data, the data may be selected by some process (e.g., women not being admitted fairly; Monty knowing which door has the money behind it), and this can induce v-structures. This results in samples that are conditioned on a common effect where the common effect is the bottom of the "V" letter in V-structure.

3.2 Importance of causal thinking

There are several benefits to causality.

1. **Active manipulation.** We have reviewed 3 ways to produce dependencies, and when there is dependence, we need to know what produced it in order to effectively manipulate and control our environments (e.g., for deciding on a treatment for kidney stones). Anytime we want to make a recommendation for taking a given action based on data, we need causality.

2. **Generalization.** Even if we don't care about manipulation, and only care about prediction, only using correlations assumes that the world is stationary. But we need causality for making predictions in the nonstationary case, where distributions change, such as if we go to a new environment (in the real world, almost everything can change, and we can never avoid new situations). For example, given a model trained in a fixed environment, we can use a person's shadow to simultaneously determine their height and the location of a light source. But in a new environment, the joint distribution of the light source's location, the person's height, and the shadow length would be different, even though the underlying physics are the same.

3. **Information Propagation and Modular Representation** Knowing the causal structure provides us with a scheme of the flow of information. This in turn will provide us with a more compact representation of relationship between variables. This modularity is indeed what is used in the case of factorizing a joint distribution with the goal of compact representation and minimal computation.

[2] `https://en.wikipedia.org/wiki/Monty_Hall_problem`

Notice that this modular representation also helps us in the context of transfer learning; if $X \to Y$ (read "X causes Y") then in a way we are assuming $p(Y|X)$ does not share information with $p(X)$ as otherwise the representation is not as compact as possible. And if these two are independent and the mechanism influencing Y through X does not change we can even infer $p(X, Y)$ when the distribution of X changes (transfer learning). For more on the latter see [7].

4. **Creativity.** Creativity requires two capacities: first, the capacity for asking "what-if" questions, known as *counterfactual reasoning* in the causal inference literature, where we imagine virtual outcomes of another world, and second, the capacity to integrate existing knowledge and information to determine how to achieve these imagined outcomes. So in a way, we need causal thinking to be creative!

4 Causal graphical representations

What's the current way for us to find causal information? First is observing the order in which things happen. For example, we know rain causes wet leaves. Usually we have temporal information; but often, the low resolution of data means we don't know which of two things happened first. However, you can imagine the following: if I have a way to *force* rain, I can see if wet leaves happen. We can imagine the effect of those changes, and then identify causal information. In a way we simulate different scenarios in our minds.

Another example: hot weather and sales of ice cream. I can find a way to increase sales of ice cream, and the weather will not change. This is how I know that heat causes ice cream sales, and not the other way around.

One other example: a person goes to work by bus in the morning. The bus coming and man leaving home seem to be related. Is there a direct causal relationship? If the man doesn't know when the bus is coming, no. But there's still a strong coincidence. Suppose guy is about to leave home, but doesn't; in that case, we would observe that the bus still comes. And, if somebody jumps in front of the bus, (hopefully) it will stop, but the man will still come to the stop. I can use those actions to identify causal relations. In this case, we are familiar with the process in the physical world, and can identify that the *timetable* is the common cause. If we change the timetable, both bus arrivals and the man leaving home will change.

An intervention is not just a change: it is much, much stronger. We can use interventions to find causal relations.

Also, to intervene on X, we have to leave all other variables unchanged. Only then can we see the effect of a particular change. Thus, interventions are defined in terms of holding everything constant but the intervention.

You can see how difficult this is to carry out: how do you change only one variable without affecting anything else?

People believe intervention and causality are circular; if we know how to apply interventions (how to intervene on X leaving all other variables unchanged except through X), we know the causality. And if we know the causality, we know how to intervene. So, understanding interventions is conceptually very important.

4.1 Causal DAGs and *do*-calculus

What is a causal DAG? First of all, it is a DAG (later, we will seem that it is not necessary to assume that the directed graph is acyclic, but for now we take DAGs). Suppose I have a causal representation:

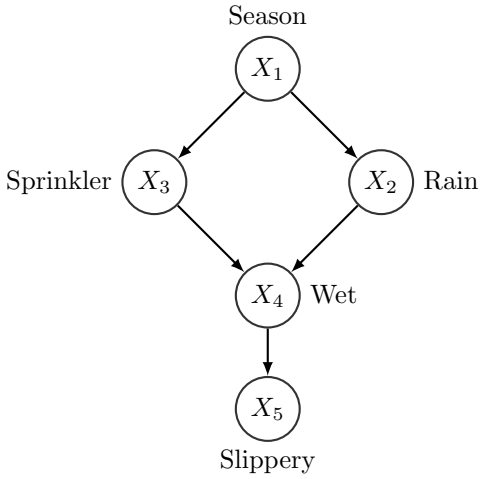

Figure 9: Causal DAG explaining the causes of slippery floor before intervention.

Then, how can I represent an intervention, for example on X_3? Set it to some particular value, in which case I don't care what causes it. When you intervene on X_3, X_1 no longer causes X_3. I can represent the causal intervention as such:

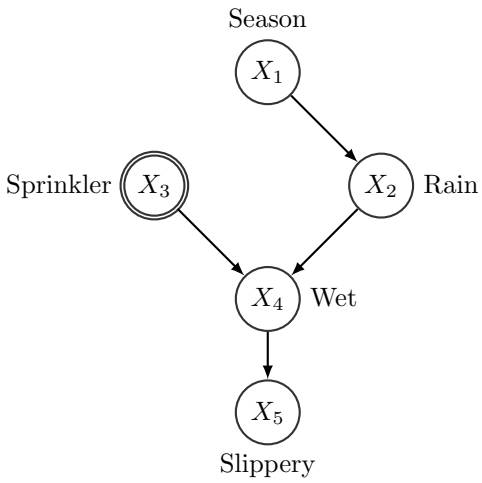

Figure 10: Causal DAG explaining the causes of slippery floor after intervention on X_3.

If this is the intervention, there is no reason for other things in the system to change, we can still make use of previous independence and conditional independent relations. We only change this random variable locally: other dependent relations will not be changed by this intervention. This is indeed the *do*-calculus we discussed in the introduction. So the initial factorization of the graph which was

$$p(X_1, X_2, X_3, X_4, X_5) = p(X_1)p(X_3|X_1)p(X_2|X_1)p(X_4|X_3, X_2)p(X_5|X_4)$$

for Figure 9 will turn to

$$p(X_1, X_2, X_3, X_4|do(X_3 = 1)) = p(X_1)p(X_2|X_1)p(X_4|X_3 = 1, X_2)p(X_5|X_4)$$

for Figure 10 which is not equivalent to the observational distribution of conditioning on $X_3 = 1$:

$$p(X_1, X_2, X_3, X_4|X_3 = 1) = \frac{p(X_1)p(X_3 = 1|X_1)p(X_2|X_1)p(X_4|X_3 = 1, X_2)p(X_5|X_4)}{p(X_3 = 1)}$$

In the world, if subsystems are independent, we can deal with each separately. Just like airplanes, where we deal with and engineer each subsystem separately. We discussed this in section 3.2.

Usually on a DAG, it is not possible to test an intervention, because how do we know the distribution of the intervention? But if a DAG is causal, we can derive interesting things using do-calculus.

4.2 Causal DAGs

A DAG G is a causal DAG if, for $P_X(V)$ is the distribution of V resulting from intervention $do(X = x)$,

1. $P_X(V)$ is Markov relative to G;
2. $P_X(V_i = v_i) = 1 for all V_i \in X$ and v_i is consistent with $X = x$; and
3. $P_X(V_i|PA_i) = P(V_i|PA_i)$ for all $V_i \notin X$, i.e., $P(V_i|PA_i)$ remains invariant to interventions not involving V_i.

4.3 Conditioning, manipulation, and counterfactual thinking

Using do-calculus on a causal DAG, we can ask three types of questions:

1. **Prediction.** Would the pavement be slipper if we *find* the sprinkler off?

$$P(Slippery|Sprinkler = off)$$

2. **Intervention.** Would the pavement be slippery if we *make sure* that the sprinkler is off?

$$P(Slippery|do(Sprinkler = off))$$

3. **Counterfactual.** Would the pavement be slippery *had* the sprinkler been off, *given that the pavement is in fact not slippery and the sprinkler is on*?

$$P(Slippery_{Sprinkler=off}|Sprinkler = on, Slippery = no)$$

5 Identification of causal effects

Let's return to the example of the kidney stone treatments (fig. 11).

We are interested in the effect of choice of treatment on recovery. This, based on what we defined so far means we are interested in $P(Recovery = 1|do(Treatment) = A)$ versus $P(Recovery = 1|do(Treatment) = B)$. But this study (and the table of outcomes) was based on observational data, not an experiment: thus, when we sum over the two stone size conditions, the column marginals are $P(Recovery|Treatment)$, but there is a latent confounder: the fact that doctors favored applying Treatment B in less severe cases, which led to different numbers in each group.

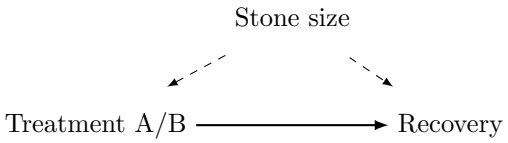

Figure 11: Again, the DAG explaining the causal relation between the variables of kidney stone problem.

Thus, the knowledge encoded in the observed distribution of $P(Recovery|Treatment)$ is not enough to identify the effect of applying one treatment versus the other, which is $P(Recovery = 1|do(Treatment))$.

More precisely, observationally (conditioning only), we have:

$$P(Recovery|Treatment) = \sum_{Stone\_size} P(Recovery|Treatment, Stone\_size)P(Stone\_size|Treatment)$$

but with experimental manipulation, we have:

$$P(Recovery|do(Treatment)) = \sum_{Stone\_size} P(Recovery|Treatment, Stone\_size)P(Stone\_size)$$

So, we break the dependence between the kidney stone size and the type of treatment. This is visually represented in figure (12).

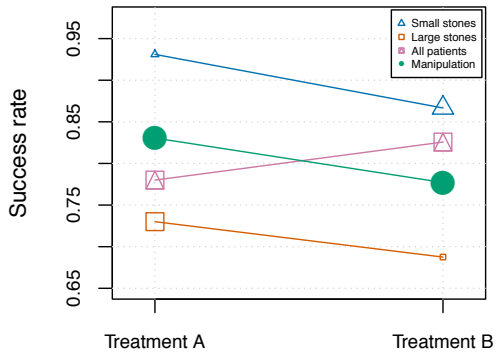

Figure 12: Comparison of total observational outcomes, outcomes conditioned on stone size, and the outcomes we would observe under proper manipulation.

In this specific case, researchers were able to identify a confounder of stone size for treatment efficacy (a confounder that led to a case of Simpson's paradox). But in general, we may not know of the existence of such confounders: they may be latent, In that case, observational knowledge is not enough to identify the distribution we want to know about.

For this reason, randomized controlled experiments (RCTs) are the "golden standard" of identifying causal effects. In RCTs, all factors that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the manipulation. In the kidney stone example, random assignment to a treatment condition (regardless of the kidney stone size) would, in expectation, lead to the number of people with small kidney stones being in treatment condition A being equal to their proportion of the overall population. From this, there would not be the confounder treatment decisions being made on the basis of stone size.

But, such as in the case of smoking causing lung cancer, RCTs may be infeasible or impossible for reasons of ethics, cost, or logistics.[3]

However, under certain conditions, even when latent variables exist one can identify these causal effects. Back-door criterion and front-door criterion are two of those two conditions.

Formally, the causal effect of X on Y is **identifiable** from a graph G if the quantity $p(y|do(x))$ can be computed uniquely from any positive probability of the observed variables—that is, for every pair of models M_1 and M_2 with $P_{M_1}(v) = P_{M_2}(v) > 0$ and $G(M_1) = G(M_2)$, $P_{M_1}(y|do(x)) = P_{M_1}(y|do(x))$.

Informally, this is asking, given the same graph and the same distribution of observations, can we find a unique representation of the causal effect? For example, in the example of treatments for kidney stones, the causal effect is *not* identifiable from observational data. Information about the causal mechanism is missing, because the same joint distribution of treatment and recovery can correspond to different causal mechanisms.

There are a number of **criteria** for if the causal effect is identifiable. Some are intuitive: the **back-door criterion** requires that a variable of interest, Y has no descendants in a set of possibly causal variables X, and that this set X blocks every path between the variable of interest Y and other variables. Intuitively, this mean that if there is a confounder, the effect of X on Y cannot be identified.

There are other criteria, that people have tried to unify. See slides for the **front-door criterion**.

6 Cycles

Feedback is when two variables affect one another over time, as in figure (13).

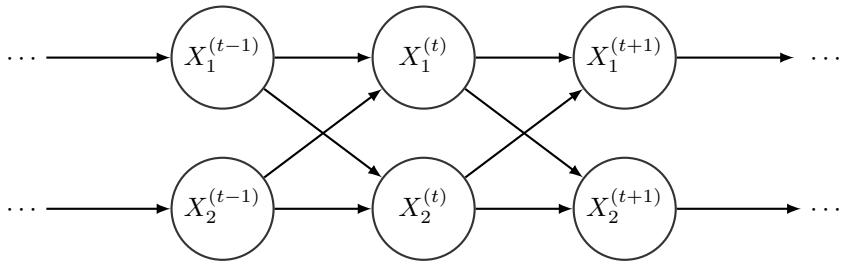

Figure 13: Feedback between X_1 and X_2.

If we represent time, there is no cycle. The "cycle" exists only conceptually if we consider the overall relationship of X_1 and X_2 (fig. 14).[4]

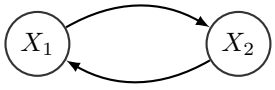

Figure 14: The representation of the "cycle" that conceptually exists between X_1 and X_2.

[3]Which again, Fisher used to argue against any evidence that smoking caused cancer!

[4]Scribe's note: other fields have names like "reciprocal determinism" and "nonlinear causality" for cycles, not necessarily thinking of processes as inevitably embedded in time.

References

[1] C. R. Charig, D. R. Webb, S. R. Payne, and J. E. A. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clinical Research Edition)*, 292(6524):879–882, 1986.

[2] Steven A. Julious and Mark A. Mullee. Confounding and Simpson's paradox. *BMJ*, 309(6967):1480–1481, 1994.

[3] Franz H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine*, 367(16):1562–1564, 2012.

[4] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[5] Scott Sinnett and Alan Kingstone. A preliminary investigation regarding the effect of tennis grunting: does white noise during a tennis shot have a negative impact on shot perception? *PloS One*, 5(10):e13148, 2010.

[6] Thomas Talhelm, X. Zhang, Shigehiro Oishi, Chen Shimin, D. Duan, Xuezhao Lan, and Shinobu Kitayama. Large-scale psychological differences within china explained by rice versus wheat agriculture. *Science*, 344(6184):603–608, 2014.

[7] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3150–3157, 2015.