Why did the distribution change?

Kailash Budhathoki **Dominik Janzing** **Patrick Blöbaum** **Hoiyi Ng**
Amazon
{kaibud, janzind, bloebp, nghoiyi}@amazon.com

Abstract

We describe a formal approach based on graphical causal models to identify the "root causes" of the change in the probability distribution of variables. After factorizing the joint distribution into conditional distributions of each variable, given its parents (the "causal mechanisms"), we attribute the change to changes of these causal mechanisms. This attribution analysis accounts for the fact that mechanisms often change independently and sometimes only some of them change. Through simulations, we study the performance of our distribution change attribution proposal. We then present a real-world case study identifying the drivers of the difference in the income distribution between men and women.

1 Introduction

Changes to a probability distribution are common in many real-world domains that are part of a changing environment. For example, during COVID-19, most retailers most likely observed a shift in the distribution of their inventory level of products— as certain products have unusually high demand (e.g., masks and hand sanitizers) while supply for some other products are limited due to suppliers suspending manufacturing. To be able to effectively respond to (either proactively or retrospectively) similar situations, it is not only important to identify if and where the distribution changed, but also know *why* the distribution changed.

In recent years, several techniques have been developed to either automatically detect changes in the underlying distribution from a sequence of observations (Pollak, 1985; Kifer et al., 2004), or determine if two data samples come from the same distribution (Chakravarti et al., 1967; Scholz and Stephens, 1987; Snedecor and Cochran, 1989; Gretton et al., 2012). However, a formal way to identify the "root causes" of the distribution change seems to be missing.

In this work, we consider a system of n variables $X_1, \ldots, X_n$. In a supply chain, for instance, these variables can represent different business metrics, such as demand forecast, labour cost, shipment cost, and inventory level to name a few. Typical *causal* questions on these variables are interventional in nature, e.g. "What would be the impact on X_k if we were to intervene on X_j?" Here we are interested in a slightly different question, namely "Which mechanisms are responsible for the change in the joint distribution $P_{X_1,\ldots,X_n}$, or the marginal distribution P_{X_k} of one of the variables X_k?" For example, in a supply chain, we might be interested in understanding the drivers of week-over-week changes in the distribution of inventory level (or a summary statistic, such as its mean across different products). To this end, we build upon graphical causal models (Pearl, 2009).

Given a causal graph G of variables $X_1, \ldots, X_n$, assuming the causal Markov condition (Spirtes et al., 2000), we can factorise the joint distribution into causal conditionals, i.e.

$$P_{X_1,\ldots,X_n} = \prod_{j=1}^{n} P_{X_j|PA_j},$$

where $P_{X_j|PA_j}$ denotes the causal mechanism of variable X_j given its direct parents PA_j in the causal graph. Each parent-child relationship captured by $P_{X_j|PA_j}$ represents an autonomous physical mechanism—we can change one such relationship *without* affecting the others.[1] Thus it is plausible to attribute any change in the joint distribution or the marginal distribution of some target variable to the change in some of the causal mechanisms. Based on this insight, we develop a formal approach to define the quantitative contribution of each mechanism to the overall change. In particular, we use the Shapley value concept (Shapley, 1953) from cooperative game theory to cope with the fact that the *impact* of changing a mechanism depends on which other mechanisms have been changed already.

The paper is structured as follows. In Section 2, we formalise changes to causal mechanisms. Section 3 presents

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] This idea of autonomy of mechanisms has a long history, see Pearl (2009, Section 1.3) and Peters et al. (2017, Section 2.2) for historical notes.

a proposal to attribute the change in the joint distribution. In Section 4, we describe a formal method to attribute the change in the marginal distribution to causal mechanisms. Section 5 discusses the practical implications of applying these attribution proposals. In Section 7, we report results from simulations and present a case study in identifying the drivers of difference in the income distribution between men and women. Finally, we conclude in Section 8.

2 Causal model and mechanism changes

We consider probabilistic causal models that incorporate probability to infer causal relationships between variables. Suppose that we have a collection of n random variables $(X_1,\dots,X_n) =: \mathbf{X}$. The underlying causal graph G of these variables is a directed acyclic graph in which a directed edge from X_i to X_j indicates that X_i causes X_j directly. A joint distribution $P_{\mathbf{X}}$ is said to be compatible with the causal graph G, if $P_{\mathbf{X}}$ can be generated following the edges in G. More formally, $P_{\mathbf{X}}$ is compatible to G if

$$P_{\mathbf{X}} = \prod_{j=1}^{n} P_{X_j|PA_j} \,. \tag{1}$$

Definition 1 (Probabilistic Causal Model) *A probabilistic causal model is a pair $\mathfrak{C} \coloneqq \langle G, P_{\boldsymbol{X}} \rangle$ that consists of a causal graph G, and a joint distribution $P_{\boldsymbol{X}}$ over the variables in G that is compatible with G.*

Given a probabilistic causal model, the causal Markov assumption (Spirtes et al., 2000) allows us to factorise the joint distribution $P_{\mathbf{X}}$ into causal mechanisms $P_{X_j|PA_j}$ at each node X_j. Each causal mechanism $P_{X_j|PA_j}$ remains invariant to interventions (external influences) in other variables. With this, we can formally define what causal mechanism changes entail.

Definition 2 (Mechanism Changes) *Mechanism changes to a causal model $\mathfrak{C} \coloneqq \langle G, P_{\boldsymbol{X}} \rangle$ on a subset of variables $\boldsymbol{X}_T$ indexed by a change set $T \subseteq \{1,\dots,n\}$ transform $\mathfrak{C}$ into $\mathfrak{C}_T \coloneqq \langle G, P^T_{\boldsymbol{X}} \rangle$, where*

$$P^T_{\boldsymbol{X}} = \prod_{j \in T} \tilde{P}_{X_j|PA_j} \prod_{j \notin T} P_{X_j|PA_j}$$

is a new joint distribution obtained by replacing "old" causal mechanism $P_{X_j|PA_j}$ at each node X_j, where $j \in T$, with the "new" causal mechanism $\tilde{P}_{X_j|PA_j}$.

The following example illustrates the idea above in a formal setting where causal relationships between variables are represented in terms of structural equations (Pearl, 2009).

Example 1 *Consider a causal model consisting of two variables, i.e. $\mathfrak{C} = \langle X_1 \to X_2, P_{X_1,X_2} \rangle$, induced by the structural equations $X_1 \coloneqq N_1$ and $X_2 \coloneqq 2X_1 + N_2$, where the independent unobserved noise terms $N_j \sim \mathcal{N}(0,1)$ are distributed according to a standard Normal distribution. A typical structure preserving intervention (Eberhardt and Scheines, 2007) changes either the parent-child functional relationship, or the distribution of unobserved noise term. Consider an intervention that changes the relationship between X_1 and X_2 from the linear function to a non-linear function represented by the new structural assignment $X_2 \coloneqq X_1^3 + N_2$. Whereas this changes the causal mechanism of X_2 from $P_{X_2|X_1}$ to $\tilde{P}_{X_2|X_1}$, the underlying causal graph remains the same. As such, we have a new causal model $\mathfrak{C}_{\{2\}} = \langle X_1 \to X_2, \tilde{P}_{X_1,X_2} \rangle$, where $P^{\{2\}}_{X_1,X_2} = P_{X_1} \tilde{P}_{X_2|X_1}$.*

The "new" joint distribution $P^T_{\mathbf{X}}$ can also be seen as the post-intervention joint distribution $P^{do(T)}_{\mathbf{X}}$ where $do(T)$ represents mechanism changes through stochastic intervention at each node X_j indexed by T (Correa and Bareinboim, 2020). In particular, we only consider the cases where the change in the joint distribution is a result of systematic replacements of independent physical mechanisms. Other cases, e.g. adversarial perturbation, can also change the joint distribution arbitrarily, which we do not consider here. Moreover, we consider the problem setting where the joint distribution changes, but *not* the causal graph.

3 Why did the joint distribution change?

As the joint distribution $P_{X_1,\dots,X_n}$ is a composition of independent causal mechanisms $P_{X_j|PA_j}$, it is plausible to attribute any change in the joint distribution to the change in some of the causal mechanisms. We would like to compute the contribution of each node—potentially due to the change in its causal mechanism—to the change in the joint distribution.

To this end, first we need a measure that quantifies the change in the joint distribution. A natural choice for quantifying the change in the joint distribution are the divergence measures as they measure the "distance" between two probability distributions. In this work, we consider the Kullback-Leibler (KL) divergence (also called relative entropy) (Cover and Thomas, 2006). Let P and Q be two probability distributions of a discrete random variable defined on the same probability space $\mathscr{X}$. Then the KL divergence from Q to P is defined as

$$D(P \,||\, Q) \coloneqq \sum_{x \in \mathscr{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

If P and Q are the distributions of a continuous random variable, then the summation is replaced by an integral. The KL divergence is particularly suitable for our purpose as it is additive for the independent compositions of the joint distribution. That is, by generalising the chain rule to more than two variables using the causal Markov condition, we get an additive decomposition of the KL divergence from the joint distribution $P_{\mathbf{X}}$ to $\tilde{P}_{\mathbf{X}}$ (Cover and Thomas, 2006,

Theorem 2.5.3) :

$$D(\tilde{P}_{\mathbf{X}} \,||\, P_{\mathbf{X}}) = \sum_{j=1}^{n} D(\tilde{P}_{X_j|PA_j} \,||\, P_{X_j|PA_j})$$

Thus, the contribution of each node X_j to the KL divergence from the joint distribution $P_{\mathbf{X}}$ to $\tilde{P}_{\mathbf{X}}$ is the KL divergence from its causal mechanism $P_{X_j|PA_j}$ to $\tilde{P}_{X_j|PA_j}$. In other words, each node—due to the change in its causal mechanism—contributes independently to the KL divergence from the joint distribution $P_{\mathbf{X}}$ to $\tilde{P}_{\mathbf{X}}$. The lemma below formalises this observation.

Definition 3 *Suppose that the causal mechanism of a node X_j changes from $P_{X_j|PA_j}$ to $\tilde{P}_{X_j|PA_j}$. Then the contribution of a node X_j to the KL divergence from the joint distribution $P_{\mathbf{X}}$ to $\tilde{P}_{\mathbf{X}}$ is $D(\tilde{P}_{X_j|PA_j} \,||\, P_{X_j|PA_j})$.*

The KL divergence is always non-negative. That is, for any P and Q, it holds that $D(P \,||\, Q) \geq 0$. Therefore, the contribution of a node X_j to the change in the joint distribution, measured in terms of the KL divergence, *cannot* be negative. If the causal mechanism of a variable did *not* change, then its contribution will be zero.

We should add a remark, however, that the KL divergence between conditionals also depends on the distribution of parents, not only the conditionals. Formally, the KL divergence from $P_{X_j|PA_j}$ to $\tilde{P}_{X_j|PA_j}$ is defined as

$$D(\tilde{P}_{X_j|PA_j} \,||\, P_{X_j|PA_j}) := \mathbb{E}_{PA_j \sim \tilde{P}_{PA_j}} \left[D(\tilde{P}_{X_j|pa_j} \,||\, P_{X_j|pa_j}) \right].$$

This is not really problematic, however, as the marginal distribution $\tilde{P}_{PA_j}$ of the parents is only used for averaging the KL divergences $D(\tilde{P}_{X_j|pa_j} \,||\, P_{X_j|pa_j})$. As such, each node X_j uses the corresponding parent-distribution $\tilde{P}_{PA_j}$ from the same joint distribution $\tilde{P}_{\mathbf{X}}$.

Estimating KL divergence in high-dimensional setting is a challenging problem. For some parametric families, such as the exponential family of distributions,[2] however, closed-form expressions exist for computing KL divergence. If a non-parametric estimator is desired, then we can use the k-nearest-neighbour-based estimator of KL divergence (Wang et al., 2009) that is asymptotically unbiased and mean-square consistent assuming i.i.d. samples.

One may also wonder whether the asymmetry of KL divergence creates a non-intuitive interpretation—as the impact of each mechanism change differs according to the direction of comparison. From an inferential standpoint, however, one direction seems preferable. When a distribution changes in time, there is a natural inferential asymmetry since one would rather consider the likelihood of new data with respect to the old model than vice versa. However, the main reason to choose KL divergence is that it nicely decomposes additively. Admittedly, this additive decomposition is a bit spoiled because we need a reference distribution to weigh the change of the conditionals—but this seems like a problem that is hard if not impossible to avoid.

In practice, often it is of interest to understand why the *marginal* distribution of one *target* variable changed, instead of the change in the joint distribution of all variables. In the next section, using a concept from game theory, we formalise how to attribute the change in the marginal distribution of a target variable to each node in the causal graph.

4 Why did the marginal distribution change?

Suppose that the marginal distribution of a target variable X_k changes—from P_{X_k} to $\tilde{P}_{X_k}$. The causal Markov condition allows us to compute the marginal distribution of any variable in the causal graph by marginalising (summing) over all independent causal mechanisms excluding that of the variable itself. Formally, given a causal model $\mathfrak{C} = \langle G, P_{\mathbf{X}} \rangle$, the marginal distribution P_{X_k} of the variable X_k can be computed by first factorising the joint distribution using the causal Markov condition, and then marginalising over all other variables, i.e.

$$\begin{aligned} P_{X_k} &= \sum_{x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n} P_{X_1,\ldots,X_n} \\ &= \sum_{x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n} \prod_{j=1}^{n} P_{X_j|PA_j} \end{aligned}$$

The change in the marginal distribution of X_k given the change set T is then given by

$$P^{T}_{X_k} = = \sum_{x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n} \prod_{j \in T} \tilde{P}_{X_j|PA_j} \prod_{j \notin T} P_{X_j|PA_j}.$$

It is therefore reasonable to assume that any change in the marginal distribution P_{X_k} of the variable X_k is most likely due to the change in some of the causal mechanisms $P_{X_j|PA_j}$. Unlike in case of the joint distribution change, however, the additive property of KL divergence cannot be leveraged directly for the attribution here—due to the marginalisation.

A natural way to compute the contribution of each node to the change in the marginal distribution of the target, from P_{X_k} to $\tilde{P}_{X_k}$, is then as follows: replace each "old" mechanism $P_{X_j|PA_j}$ by the "new" mechanism $\tilde{P}_{X_j|PA_j}$ in succession. Each replacement changes the marginal distribution of the target X_k, which can be used to compute the contribution of the corresponding node. The amount of change, however, depends on the causal mechanisms that have been already replaced. In other words, the contribution of a node X_j to the change in the marginal distribution of the target X_k depends on the order in which we replace the causal mechanisms—the resulting attribution procedure is hence in danger of becoming arbitrary.

[2] The exponential family of distributions includes the Gaussian, Poison, Binomial, Multinomial, and Beta, as well as many others.

4.1 Shapley Values

The Shapley value (Shapley, 1953) from cooperative game theory provides a principled approach to mitigate the arbitrariness in the attribution procedure, due to the dependence on the ordering of replacements. In particular, it removes the arbitrariness by symmetrizing over all orderings. We briefly summarise the Shapley value here.

Let $N \coloneqq \{1,\dots,n\}$ be a set of n players and $\nu : 2^N \to \mathbb{R}$ be a set function that associates a real-valued payoff to a coalition of players $T \subseteq N$ with $\nu(\emptyset) = 0$, where $\emptyset$ denotes an empty set. We assume that players will cooperate to form a grand coalition N. The goal is then to "fairly" assign the resulting payoff $\nu(N)$ to each player j in N.

Let $\sigma : N \to N$ denote a permutation of players N. All permutations of the set N with n elements form a symmetric group S_n. Suppose that each player enters into a coalition one by one in the ordering $\sigma(1), \sigma(2), \dots, \sigma(n)$ to eventually form a grand coalition. Let $N_{prec}(j)$ denote the set of players that precede player j in the ordering, i.e. $N_{prec}(j) \coloneqq \{i \in N \mid \sigma^{-1}(i) < \sigma^{-1}(j)\}$. Note that $\sigma^{-1}(i)$ gives player i's position. Then the change in the payoff of a coalition when a player joins the coalition is the marginal contribution of the player to the coalition,

$$C_\nu\left(j \mid N_{prec}(j)\right) \coloneqq \nu\big(N_{prec}(j) \cup \{j\}\big) - \nu\big(N_{prec}(j)\big) \,.$$

The marginal contributions of all players will then sum up to the payoff of the grand coalition $\nu(N)$, i.e.

$$\sum_{j=1}^{n} C_\nu\left(j \mid N_{prec}(j)\right) = \nu(N) \,. \tag{2}$$

Unfortunately, the marginal contribution of a player j depends on the order σ in which players enter into a coalition. To remove the dependence on the ordering σ, Shapley's solution (Shapley, 1953) is to assign each player $j \in N$ its average marginal contribution over all permutations, i.e.

$$\begin{aligned}
\phi_j(\nu) &\coloneqq \frac{1}{n!} \sum_{\sigma \in S_n} \nu\big(N_{prec}(j) \cup \{j\}\big) - \nu\big(N_{prec}(j)\big) \\
&= \sum_{T \subset N \setminus \{j\}} \frac{|T|!(n-|T|-1)!}{n!} \left(\nu(T \cup \{j\}) - \nu(T)\right) \\
&= \sum_{T \subset N \setminus \{j\}} \frac{1}{n\binom{n-1}{|T|}} \left(\nu(T \cup \{j\}) - \nu(T)\right) \,,
\end{aligned}$$

where the second line follows from counting the number of permutations (out of $n!$) where $N_{prec}(j) = T$. The Shapley value is also desirable because it gives a unique solution to the following axioms that capture the notion of fairness.

Efficiency The Shapley values of all players sum up to the payoff of the grand coalition. That is, $\sum_{j=1}^{n} \phi_j(\nu) = \nu(N)$ for any set function ν.

Symmetry If two players contribute the same amount to every coalition of other players, then their Shapley values are equal. Formally, for two players i and j, if $\nu(T \cup \{i\}) = \nu(T \cup \{j\})$ for all coalitions $T \subseteq N \setminus \{i, j\}$, then we have $\phi_i(\nu) = \phi_j(\nu)$.

Null Player A player that does not contribute to any coalition will get a zero Shapley value. Formally, for a player j, if $\nu(T \cup \{j\}) = \nu(T)$ for all coalitions $T \subset N \setminus \{j\}$, then $\phi_j(\nu) = 0$.

Additivity The Shapley value computed from the sum of two set functions is the same as the sum of the Shapley values computed using individual set functions. That is, for any two set functions ν_1 and ν_2, we have $\phi_j(\nu_1 + \nu_2) = \phi_j(\nu_1) + \phi_j(\nu_2)$.

Computing the Shapley value for a player has the worst-case time-complexity of $\mathcal{O}(2^n)$. However, efficient approximations exist (Lundberg and Lee, 2017).

4.2 Attributing Marginal Distribution Change

By slightly abusing the notation, we use the index set $\{1,\dots,n\} =: N$ to refer to the corresponding variables $X_1,\dots,X_n$. Then any coalition $T \subset N$ represents the change set for mechanism changes to the "old" causal model. Let $P^T_{X_k}$ denote the marginal distribution of X_k obtained by marginalising the joint distribution $P^T_{\mathbf{X}}$ in the new causal model $\mathfrak{C}_T \coloneqq \langle G, P^T_{\mathbf{X}} \rangle$. Naturally, for $T = \emptyset$, the causal model does not change, i.e. $\mathfrak{C}_T = \mathfrak{C}$ for $T = \emptyset$.

First consider a scenario where the change in the marginal distribution of the target X_k is quantified by the KL divergence. The quantity that we want to attribute to each node is then the KL divergence $D(\tilde{P}_{X_k} \,||\, P_{X_k})$. To this end, we define the marginal contribution of a node X_j given a change set T as

$$C_D\left(j \mid T\right) \coloneqq D(P^{T \cup \{j\}}_{X_k} \,||\, P_{X_k}) - D(P^{T}_{X_k} \,||\, P_{X_k}) \,.$$

In other words, the marginal contribution quantifies how much replacing the causal mechanism of variable X_j contributes to the KL divergence $D(\tilde{P}_{X_k} \,||\, P_{X_k})$, given that we have already replaced the causal mechanisms of variables in the change set T. Then the Shapley value of the variable X_j is

$$\phi_j(D) \coloneqq \sum_{T \subset N \setminus \{j\}} \frac{1}{n\binom{n-1}{|T|}} C_D\left(j \mid T\right), \tag{3}$$

which gives the "fair" contribution of X_j to the change in the marginal distribution of the target X_k measured in terms of $D(\tilde{P}_{X_k} \,||\, P_{X_k})$. Note that the Shapley value contribution $\phi_j(D)$ can be negative. This is completely reasonable as the marginal contribution $C_D\left(j \mid T\right)$ can be negative; replacing the causal mechanism of a variable can bring the

"new" marginal distribution of the target closer to the "old" marginal distribution P_{X_k}, thereby lowering the KL divergence relative to not replacing it.

Due to the efficiency property, the Shapley values of all variables will sum up to the KL divergence of the marginal distribution from P_{X_k} to $\tilde{P}_{X_k}$, i.e.

$$\sum_{j=1}^{n} \phi_j(D) = D(\tilde{P}_{X_k} \,||\, P_{X_k}) \,.$$

Note that the equality above may not hold strictly when we approximate the Shapley values for players. For the worst case analysis on the approximation error of Shapley values, we refer the interested reader to Charnes et al. (1988); Fatima et al. (2008).

Instead of the overall change in the marginal distribution as measured by the KL divergence, we might be interested in the change in some of its property or summary (e.g. mean, median, variance, skew). For instance, it is not uncommon for a retailer to ask, "Why did the mean inventory level go down?", as maintaining an inventory requires upfront investment. Amongst many, that could be due to the change in the distribution of the demand forecast, or some changes in the algorithms (mathematical functions). The definition below provides a rather general way to attribute marginal distribution change.

Definition 4 (Marginal Distribution Change Attribution) *Let ψ denote any functional defined on the marginal distribution. Given a change set T, the marginal contribution of a node X_j to the change in the functional of the marginal distribution of the target X_k, i.e. $\Delta\psi \coloneqq \psi(\tilde{P}_{X_k}) - \psi(P_{X_k})$, is*

$$C_\psi(j \mid T) \coloneqq \psi(P_{X_k}^{T\cup\{j\}}) - \psi(P_{X_k}^{T}),$$

and its Shapley value contribution to $\Delta\psi$ is given by

$$\phi_j(\psi) \coloneqq \sum_{T\subset N\setminus\{j\}} \frac{1}{n\binom{n-1}{|T|}} C_\psi(j \mid T) \,.$$

The marginal contribution $C_\psi(j \mid T)$ quantifies how much replacing the causal mechanism of X_j contributes to the change in the summary of P_{X_k}, given that we have already replaced the causal mechanisms of variables in the change set T. As the marginal contribution of a variable can be negative, the Shapley value contribution $\phi_j(\psi)$ can also be negative. In the example below, we show how to attribute the change in the mean of the target variable to each node.

Example 2 *Let $\mathbb{E}_{X_k\sim P_{X_k}}[X_k]$ denote the mean of the variable X_k under the distribution P_X. The Shapley value contribution of a node X_j to the change in the mean (due to the change in the marginal distribution) of the target node X_k is*

$$\phi_j(\mathbb{E}) \coloneqq \sum_{T\subset N\setminus\{j\}} \frac{1}{n\binom{n-1}{|T|}} C_{\mathbb{E}}(j \mid T) \,,$$

where the marginal contribution $C_{\mathbb{E}}(j \mid T)$ of X_j given a change set T is defined as

$$C_{\mathbb{E}}(j \mid T) \coloneqq \mathbb{E}_{X_k\sim P_{X_k}^{T\cup\{j\}}}[X_k] - \mathbb{E}_{X_k\sim P_{X_k}^{T}}[X_k] \,.$$

The Shapley values of all nodes sum up to the change in the mean of the target node. That is, the following holds:

$$\sum_{j=1}^{n} \phi_j(\mathbb{E}) = \mathbb{E}_{X_k\sim \tilde{P}_{X_k}}[X_k] - \mathbb{E}_{X_k\sim P_{X_k}}[X_k] \,.$$

Next we discuss practical aspects of distribution change attribution solution we have presented thus far.

5 Detecting mechanism changes

To apply our attribution methods, we require a causal graph and its causal conditionals. Existing techniques on sound and complete causal structure learning (Spirtes et al., 2000; Pearl, 2009), however, can only recover the Markov equivalence class of DAGs assuming faithfulness. With additional assumptions on the data-generating process, it is possible to recover the exact causal graph (Peters et al., 2017; van de Geer and Bühlmann, 2013). A promising approach is to combine domain knowledge with conditional independence tests (Mastakouri et al., 2019). In some cases, we can also perform controlled randomised experiments to identify the exact DAG from the Markov equivalence class (Dubois et al., 2020).

Once we have the causal graph, we can then estimate the causal conditionals directly from data using techniques from high-dimensional statistics (Bühlmann and van de Geer, 2011; Wainwright, 2019). Note that we are given two samples of the same size or different sizes (e.g. data from the week before Christmas, and data from the Christmas week). Then for each node X_j, we estimate $P_{X_j|PA_j}$ from the first sample, and $\tilde{P}_{X_j|PA_j}$ from the second sample. In the context of distribution-change attribution, however, sampling variability can lead to spurious results when we directly plug in the estimated causal conditionals. Even if two samples are drawn from the same joint distribution P_{X_j,PA_j}, we will most likely estimate two different causal conditionals (even with regularisation), because of sampling variability. Contrary to the expectation, X_j will then be attributed a non-zero value.

If we knew that the causal mechanism $P_{X_j|PA_j}$ did not change, it would then make sense to learn the causal mechanism from the combined sample, as then the quality of the learned causal conditional will also improve due to the increased sample size. Moreover, we can also directly attribute a zero contribution to the node. This raises the question, "how do we detect causal mechanism changes from two samples?" In other words, are there any statistically testable implications of causal mechanism changes?

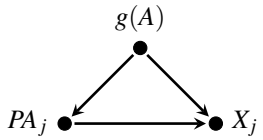

Figure 1: The assumed causal graph to detect mechanism changes for a parent-child relationship.

There exists a large body of work, e.g. Chakravarti et al. (1967); Scholz and Stephens (1987); Snedecor and Cochran (1989); Gretton et al. (2012), on statistical hypothesis test to determine whether the difference between the two distributions is statistically significant from their two samples, each drawn from a separate distribution. Those are not applicable in our setting as we are interested in the difference between the *conditional* distributions, not the marginals. In a recent work, Huang et al. (2020) extend the PC algorithm Spirtes et al. (2000) for causal discovery from non-stationary or heterogeneous data, where one of the steps involve detecting changing causal mechanisms. Here we adapt their method.

Assume that causal mechanisms $P_{X_j|PA_j}$ can be written as functions of a time or domain index A (see Figure 1). If the causal graph is induced by a functional causal model, then the quantities such as functional models, noise levels, etc that may change over time or across domains can be written as functions of A. Under these assumptions, if the causal mechanism $P_{X_j|PA_j}$ remains the same across various values of A, then the conditional independence test $X_j \perp\!\!\!\perp A \mid PA_j$ suffices to detect changes to the causal mechanism $P_{X_j|PA_j}$.

Let D_t denote the m_t-by-n matrix containing sample from time (or domain) $t \in \{1,2\}$, and D denote the m-by-n matrix obtained by vertically concatenating D_ts, where $m = m_1 + m_2$. That is, we have

$$D := \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}$$

We can construct A directly from data. The key idea is to assign the same value of A to the units in the sample from the same time (or domain) t. For clarity, with a slight abuse of notation, we interchangeably use A for a variable as well as the data vector. Each entry a_i of m-by-1 vector A is assigned

$$a^{(i)} := \begin{cases} +1, & \text{if } i \leq m_1, \\ -1, & \text{otherwise.} \end{cases}$$

Using the columns from the combined data matrix D and index vector A, we then test if each variable X_j is conditionally independent of A given its direct parents PA_j in the causal graph, i.e. $X_j \perp\!\!\!\perp A \mid PA_j$. If the conditional independence holds, then the causal mechanism $P_{X_j|PA_j}$ did not change across various values of A. On the other hand, if X_j is dependent on A given PA_j then its causal mechanism $P_{X_j|PA_j}$ also changes with A.

Note that the functional relationships between variables X_j and time (or domain) index A is unknown. It is therefore important to use a non-parametric conditional independence test. In this work, we use kernel-based conditional independence test (Zhang et al., 2011).

6 Related Work

Path-specific effect of X and Y via path π is the degree to which an interventional-change in X would change the marginal distribution of Y if that change were to be transmitted only via π. If an indicator (or context) variable A represents samples from distributions P and $\tilde{P}$, then computing the path-specific effect of A on any node via the direct path simply measures the distance between the "old" causal mechanism and the "new" causal mechanism of the node. Arguably, we then capture the causal influence of external factors (abstracted by A) to the node (Janzing et al., 2013). To discuss the relation to the strength of causal arrows defined in Janzing et al. (2013), we need to label the two distributions by an additional variable V attaining values v and $\bar{v}$ with edges to all X_j whose mechanisms change. Their strengths would be $D(P_{X_j|PA_j} \,||\, [p(v)P_{X_j|PA_j} + p(\bar{v})P_{X_j|PA_j}])$.

Most feature-based interpretability techniques assume that features independently co-cause the target (Lundberg and Lee, 2017; Janzing et al., 2020). In particular, they consider how the marginal distribution of the target changes w.r.t. to interventional changes in the features. They cannot attribute causal mechanism changes, as one must assume that the causal mechanism of the target does not change to make them work for multiple datasets.

In causal discovery from multiple contexts (Mooij et al., 2020), they find the union of causal graphs in each dataset (or context) by jointly modelling the context variables $(A_1, \ldots, A_k)$ and observed variables $(X_1, \ldots, X_n)$. While they also work on samples from different distributions and hence might appear related, these are two different problems.

Mechanism changes can also be represented with stochastic interventions (Pearl, 2009; Correa and Bareinboim, 2020). We briefly discussed the connection in the last paragraph of Section 2. Computing the effect of a stochastic intervention, however, does not tell us how much the corresponding mechanism change contributed to a target quantity summarising distribution change.

Overall, existing methods are not suitable for distribution change attribution, where our goal is to attribute a target quantity summarising distribution change (e.g. change in mean, KL divergence) to change in causal mechanism (conditional distribution of a node given its direct causes) of each variable.

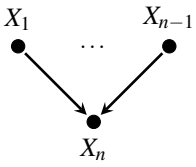

Figure 2: The causal graph used in simulations, where independent inputs $X_1, \ldots, X_{n-1}$ co-cause the target X_n.

7 Experiments

In experiments, we study the performance of our proposal for attributing the change in the marginal distribution, and its application to real-world data. In particular, through simulations, we study the performance of our attribution proposal when we learn causal mechanisms from samples w.r.t. sample size and magnitude of causal mechanism change. On real-world data, we asses whether results are sensible.

7.1 Simulations

We consider a setting where $n-1$ independent input variables $X_1, \ldots, X_{n-1}$ co-cause a target variable X_n. Their underlying causal graph is shown in Figure 2. We choose this simple, yet representative, causal graph for simulation because computing the Shapley values analytically for summaries of distributional changes is not trivial for rather complex graphs.

Let $N_w \sim \mathcal{N}(\mu_w, 1)$ denote an independent Gaussian noise with mean μ_w and unit variance for each $w \in \{1, \ldots, n\}$. Suppose that their causal graph is induced by the following structural assignments:

$$\begin{aligned} X_w &:= N_w \text{ for } w \in \{1, \ldots, n-1\} \text{ and} \\ X_n &:= X_1 + \ldots + X_{n-1} + N_n. \end{aligned}$$

We refer to the structural causal model (SCM) above by $\mathfrak{C}$. Suppose that their joint distribution $P_{X_1,\ldots,X_n}$ changes to $\tilde{P}_{X_1,\ldots,X_n}$ due to the changes in the structural assignments. In the new SCM $\tilde{\mathfrak{C}}$, we have the following assignments:

$$\begin{aligned} X_w &:= N_w + \lambda_w \text{ for } w \in \{1, \ldots, n-1\} \text{ and} \\ X_n &:= X_1 + \ldots + X_{n-1} + N_n + \lambda_n, \end{aligned}$$

where λ_w is a scalar that shifts the mean of the corresponding variable X_w for each $w \in \{1, \ldots, n\}$, and further defined as

$$\lambda_w := \begin{cases} \lambda & \text{if } S_w = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where S_w is a Bernoulli random variable with a probability of success $p = 0.5$. That is, S_w determines whether the causal mechanism of the corresponding variable X_w is potentially subject to change. If $S_w = 1$, then the value of λ subsequently dictates the *magnitude* of the change in the causal mechanism of X_w. Note that even if $S_w = 1$, the causal mechanism of corresponding variable X_w does not change if $\lambda = 0$. With rejection sampling, we ensure that at least one causal mechanism changes, i.e. $\lambda_w \neq 0$ for at least one $w \in \{1, \ldots, n\}$. This way, we can change the causal mechanism of a random subset of variables through S_w, and study the performance of our proposal w.r.t λ.

Due to changes in the causal mechanisms of variables, the marginal distribution of the target X_n also changes:

$$\begin{aligned} \mathfrak{C} &: X_n \sim \mathcal{N}(\mu_1 + \ldots + \mu_n, n) \\ \tilde{\mathfrak{C}} &: X_n \sim \mathcal{N}(\mu_1 + \ldots + \mu_n + \lambda_1 + \ldots + \lambda_n, n). \end{aligned}$$

Let P_{X_n} and $\tilde{P}_{X_n}$ denote the marginal distributions of X_n in SCMs $\mathfrak{C}$ and $\tilde{\mathfrak{C}}$ respectively. We measure the change in the marginal distribution of X_n by the difference in its mean, i.e.

$$\begin{aligned} \Delta\mathbb{E} &:= \mathbb{E}_{X_n \sim \tilde{P}_{X_n}}[X_n] - \mathbb{E}_{X_n \sim P_{X_n}}[X_n] \\ &= \lambda_1 + \ldots + \lambda_n. \end{aligned}$$

With some algebraic manipulation, we can show that the contribution of each variable X_w—due to the change in its causal mechanism—to $\Delta\mathbb{E}$ is then given by

$$\phi_w(\mathbb{E}) = \lambda_w \tag{5}$$

These closed-form expressions provide the ground truth for our evaluation. As it should, we see that the following holds:

$$\phi_1(\mathbb{E}) + \ldots + \phi_n(\mathbb{E}) = \Delta\mathbb{E}.$$

First we generate two samples of the same size from the two SCMs $\mathfrak{C}$ and $\tilde{\mathfrak{C}}$ stated above. From each sample, we learn (estimate) the SCM assuming that the causal graph is known. As the SCM has an additive unobserved noise term, it is possible to estimate both the function and the noise from data with regression. We generate a sample from the joint distribution induced by the learned SCM. Then we estimate the mean of the marginal distribution by the sample average. Finally, we compute the Shapley value of each variable X_w. To measure the quality of the estimated Shapley values against the ground truth, we use the ℓ_1 norm, otherwise known as Manhattan distance.

First we study the performance of our attribution proposal against the magnitude parameter λ. To this end, for a given value of λ, we generate 100 pair of SCMs ($\mathfrak{C}$, $\tilde{\mathfrak{C}}$) with μ_w chosen according to the Uniform distribution $\mathcal{U}(-5,5)$ for each $w \in \{1, \ldots, n\}$, where n is chosen uniformly randomly from $\{2,3,4,5\}$. From each SCM in the pair ($\mathfrak{C}$, $\tilde{\mathfrak{C}}$), we generate 100 samples, each containing 1000 observations. Note that we learn the SCM from each sample, and then estimate the Shapley values from the sample drawn from the learned SCM, which we repeat 100 times. Therefore we report the average ℓ_1 distance (with standard error) over $100 \times 100 \times 100 = 1\,000\,000$ pairs of samples in Figure 3 at

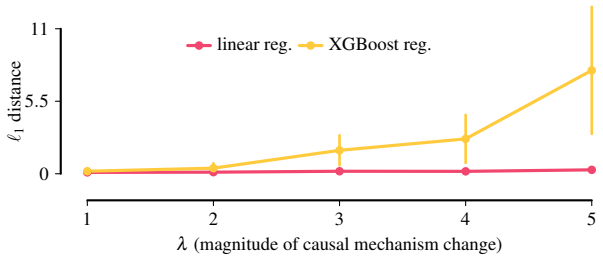

Figure 3: The ℓ_1 distance between the ground truth, and the estimated Shapley values when the underlying linear structural causal model is estimated with the linear regression model versus the graident boosted trees regression model at various values of λ. The standard error bars for the linear regression model are invisble as they are too narrow.

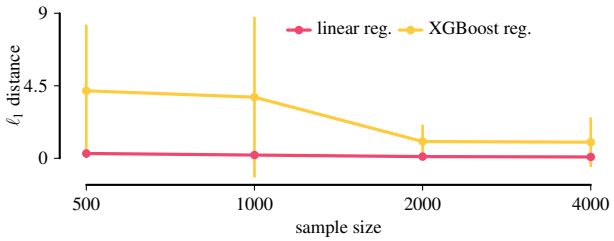

Figure 4: The ℓ_1 distance between the ground truth, and the estimated Shapley values when the underlying linear structural causal model is estimated with linear regression versus XGBoost regression at various sample sizes.

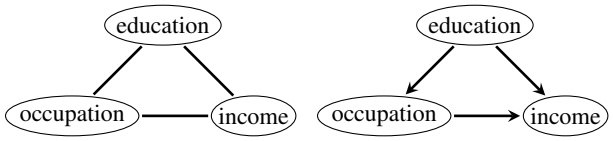

Figure 5: (left) Skeleton graph discovered using PC algorithm on a subset of variables from the Adult Census Income dataset. (right) Causal graph derived from the skeleton by orienting the undirected edges using domain knowledge.

various values of λ. With a right regression model (linear regression), the estimated Shapley values are very close to the ground truth regardless of the magnitude of causal mechanism change λ—indicated by close to zero ℓ_1 distance. With gradient boosted trees (from `xgboost` python package with default hyperparameters and 100 trees), however, the estimated Shapley values, on average, differ from the ground truth as λ increases. This is expected as inferring the right model is harder if function classes are less restricted a priori. Therefore, the Shapley values estimated using a non-linear function will deviate from the ground truth compared to that from a linear function. We also observe that the uncertainty in the Shapley value estimation increases if model estimation is not accurate. This is indicated by wide error bars for the gradient boosted trees compared to narrow error bars, that are invisible in the figure, for the linear regression.

In Figure 4, we show the result when we vary the sample size, but randomly choose the magnitude parameter λ according to $\mathscr{U}(1,5)$ for each SCM pair. We observe that the Shapley values from linear regression model is close to the ground truth even at a relatively small sample size of 500—with an average ℓ_1 distance of 0.29 and standard error of 0.21. While the performance of XGBoost regression model certainly improves with increasing sample size, its performance does not match the linear regression model.

7.2 Case Study

Next we present a case study on the Adult Census Income dataset[3] where we use our proposal to identify the drivers of the difference in the income distribution between men and women.

The dataset contains 32,561 records from the census on annual income in the United States from 1994. In addition to whether the annual income of an individual is greater than fifty thousand USD, it contains 14 other socio-economic attributes. We consider a subset of non-redundant attributes for analysis, namely education and occupation, that directly affect income as well as act a proxy of income for other attributes. After removing the rows with missing values, we end up with 30,718 rows.

As the number of variables is small, we combine causal discovery with domain knowledge to construct the causal graph. First, from the combined records of men and women, we discover the skeleton graph using the PC algorithm (Spirtes et al., 2000) with kernel-based conditional independence test (Zhang et al., 2011) at a significance level of 0.2 (see Figure 5 left). At a higher significance level, the PC algorithm gives us a denser skeleton. This way we minimise the chances of omitting dependencies. We then orient the edges in the skeleton using domain knowledge. Note that changing the occupation will "mainly" lead to the change in the annual income, not the other way around. Changing the education not only affects the occupation, but also the income (Heckman et al., 2018), e.g. passive income through smart investments, side incomes, etc. Therefore, education confounds both occupation and income. We then have a causal graph as shown in Figure 5 (right).

Since our goal is to identify the drivers of difference in the income distribution between men and women, as a sanity check, we perform a two-sample test to determine whether the income distribution is, indeed, different between men and women in the dataset. Under the null hypothesis that incomes of men and women come from the same distribution, the Kolmogorov-Smirnov two-sample test yields a p-value

[3] `http://archive.ics.uci.edu/ml/datasets/Adult`

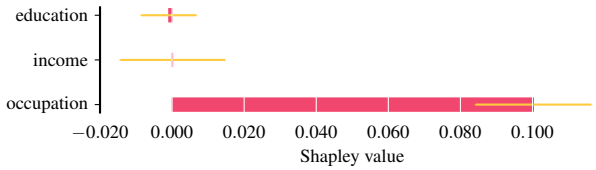

Figure 6: Shapley value contribution of each variable (due to the potential difference in its causal mechanism) to the difference in the mean annual income between men and women ($\mu_{\text{men}} - \mu_{\text{women}}$). Each horizontal bar represents the mean Shapley value contribution, and each line represents the bias-corrected and accelerated bootstrap confidence interval at a 95% confidence level, over 100 resamples. The horizontal bars of education and income are almost invisible as their mean Shapley value contributions are close to zero.

of 1.38×10^{-234}. Since the p-value is extremely low, we can safely reject the null hypothesis.

All three variables are categorical. In particular, the target variable (income) is binary (>50K?). We use the empirical distribution as the causal mechanism of the root node (education). For a non-root node (occupation and income), we learn its causal mechanism using a XGBoost classifier with 100 gradient boosted trees. To detect causal mechanism changes, we use kernel-based conditional independence test (Gretton et al., 2012) with delta kernel at a significance level of 0.05. We would like to attribute the difference in the *mean* annual income between men and women.

The result of our method is shown in Figure 6. We find that the change in the causal mechanism of occupation, $P_{\text{occupation}|\text{education}}$, is the main driver for the difference in the mean annual income between men and women. This is also often cited in public discourse as a reason why we need more women participation in labour market to empower them economically (Giuliano, 2017).[4] Educational choices of men and women differ—science-related subjects are attended mostly by men than women (Wang and Degol, 2017; Tellhed et al., 2016). In UC Berkeley gender bias study from 1975, for instance, it was found that, compared to men, women tended to apply to departments that are more crowded, less well funded, and that frequently offer poorer professional employment prospects (Bickel et al., 1975). Graduates of science-related subjects are also known to earn more than the others (Deming and Noray, 2018). Moreover, women often take non-professional responsibilities, such as parenting and caring family relatives that affect their career and earnings (Jolly et al., 2014; Budig, 2006). Therefore, income distribution of men and women differ even when they have the same level of education. As a representative case, we show the empirical conditional probability distribution of occupation given "Bachelor" educated men versus women in Figure 7, which also corroborates our attribution result. The finding that the subject of education plays a significant role on occupation and income would deserve further studies with more detailed data sets which goes beyond the scope of the present paper.

[4] A comprehensive review of literature on this topic along with data and visualisation is available at `https://ourworldindata.org/female-labor-supply`.

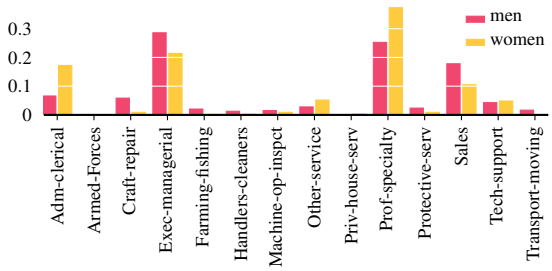

Figure 7: Conditional probability distribution of occupation given "Bachelor" educated men versus women.

8 Discussion and Conclusions

We presented a formal approach to identify the drivers of distribution change using graphical causal models. The key idea is that, given a causal graph, we can factorise the joint distribution into independent causal conditionals. Any change in the joint distribution, or marginal distribution of any target variable thereof, can then be attributed to changes in some of the causal conditionals. We illustrated our method on both simulated and real-world datasets.

In Section 5, we showed how to detect causal mechanism changes from data given the underlying causal graph using conditional independence tests. In many tasks—e.g. exploratory data analysis, designing and evaluating robust models—knowing those conditionals that change is already sufficient. One might then ask, "Why do we need to quantify the contribution from each mechanism?" Causal mechanisms always change in a system of a large number of variables that are embedded in a changing environment. Supply chain is one such example where continuous deployments of new changes in constituent subsystems (e.g. forecasting, buying) are common. It is too costly (e.g. time, personnel) to look at all variables whose mechanisms change. Quantifying how much each variable contributed to the change allows us to focus on a few most relevant variables.

Finally, our attribution proposal requires a causal graph, which may not be identifiable from observational data. If the causal graph is not identifiable, the Shapley values will not be identifiable as well. Therefore, this question on the robustness of our attribution proposal to causal graph misspecification deserves further research.

Acknowledgements

The authors thank Dr. David Afshartous for useful comments.

References

P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187 (4175):398–404, 1975.

M. J. Budig. Gender, self-employment, and earnings: The interlocking structures of family and professional status. *Gender & Society*, 20(6):725–753, 2006.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2011.

I. M. Chakravarti, R. G. Laha, and J. Roy. Handbook of methods of applied statistics, 1967.

A. Charnes, B. Golany, M. Keane, and J. Rousseau. *Extremal Principle Solutions of Games in Characteristic Function Form: Core, Chebychev and Shapley Value Generalizations*, pages 123–133. Springer Netherlands, 1988.

J. Correa and E. Bareinboim. A calculus for stochastic interventions:causal effect identification and surrogate experiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10093–10100, 2020.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

D. J. Deming and K. L. Noray. Stem careers and the changing skill requirements of work. Working Paper 25065, National Bureau of Economic Research, 2018.

J. Dubois, H. Oya, J. M. Tyszka, M. Howard, F. Eberhardt, and R. Adolphs. Causal mapping of emotion networks in the human brain: Framework and initial findings. *Neuropsychologia*, 145, 2020.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981–995, 2007.

S. S. Fatima, M. Wooldridge, and N. R. Jennings. A linear approximation method for the shapley value. *Artificial Intelligence*, 172(14):1673–1699, 2008.

P. Giuliano. Gender: An historical perspective. NBER Working Papers 23635, National Bureau of Economic Research, Inc, 2017.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

J. J. Heckman, J. E. Humphries, and G. Veramendi. Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking. *Journal of Political Economy*, 126(S1):197–246, 2018.

B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324 – 2358, 2013.

D. Janzing, L. Minorics, and P. Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 2020.

S. Jolly, K. A. Griffith, R. DeCastro, A. Stewart, and P. Ubel. Gender differences in time spent on parenting and domestic responsibilities by high-achieving young physician-researchers. *Annals of Internal Medicine*, 160(5):344–353, 2014.

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, page 180–191. VLDB Endowment, 2004.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

A. Mastakouri, B. Schölkopf, and D. Janzing. Selecting causal brain features with a single conditional independence test per feature. In *Advances in Neural Information Processing Systems 32*, pages 12532–12543. Curran Associates, Inc., 2019.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

M. Pollak. Optimal detection of a change in distribution. *Annals of Statistics*, 13(1):206–227, 1985.

F. W. Scholz and M. A. Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games (AM-28)*, 2, 1953.

G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, 8th edition, 1989.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

U. Tellhed, M. Bäckström, and F. Björklund. Will i fit in and do well? the importance of social belongingness and self-efficacy for explaining gender differences in interest in stem and heed majors. *Sex Roles*, 77, 10 2016.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

M. T. Wang and J. L. Degol. Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educ Psychol Rev*, 29(1):119–140, 2017.

Q. Wang, S. R. Kulkarni, and S. Verdu. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55 (5):2392–2405, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. pages 804–813, Corvallis, OR, USA, 2011. AUAI Press.