Operationalizing Complex Causes: A Pragmatic View of Mediation

Limor Gultchin [1 2] **David S. Watson** [3] **Matt J. Kusner** [4] **Ricardo Silva** [3 2]

Abstract

We examine the problem of causal response estimation for complex objects (e.g., text, images, genomics). In this setting, classical *atomic* interventions are often not available (e.g., changes to characters, pixels, DNA base-pairs). Instead, we only have access to indirect or *crude* interventions (e.g., enrolling in a writing program, modifying a scene, applying a gene therapy). In this work, we formalize this problem and provide an initial solution. Given a collection of candidate mediators, we propose (a) a two-step method for predicting the causal responses of crude interventions; and (b) a testing procedure to identify mediators of crude interventions. We demonstrate, on a range of simulated and real-world-inspired examples, that our approach allows us to efficiently estimate the effect of crude interventions with limited data from new treatment regimes.

1. Introduction

Understanding causal mechanisms is a primary goal of scientific inquiry and a crucial prerequisite for planning effective interventions. However, the task of isolating and quantifying treatment effects is complicated by several obstacles. Fundamental questions of identifiability (Shpitser and Pearl, 2008; Correa and Bareinboim, 2020) and transportability (Bareinboim and Pearl, 2016) pose significant challenges to practitioners across a variety of disciplines. The problems are particularly acute in high-dimensional settings where interventions are rarely of the surgical or "atomic" sort envisioned by most authors in this area. For instance, genomic data contains rich information about the pharmacodynamic impact of drug therapies on disease activity. However, careful analysis is required to detect and operationalize these sparse signals, as causal effects are not defined in terms of direct interventions on, say, individual genes, but are instead propagated from a crude treatment (drug administration) on a complex object (the human transcriptome), which affects outcomes (disease activity) through several mediating pathways. Similar complexity arises in other fields, for instance when purported causes are social constructs like "gross domestic product" (Arnold et al., 2020) or large-scale natural phenomena like "El Niño" (Chalupka et al., 2016).

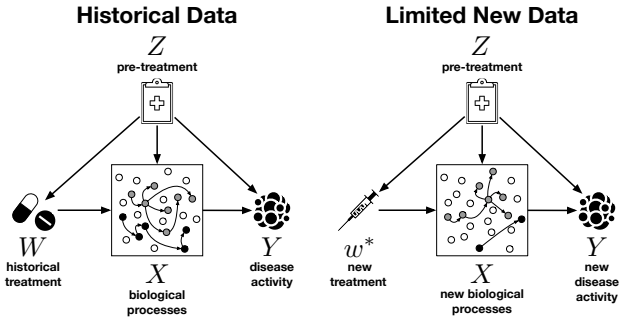

Figure 1. The complex cause problem setting. See text for details.

Despite a substantial and growing literature on causal inference (see Sect. 2), existing theory largely fails to accommodate complex systems where the putative causes X of an outcome of interest Y have many internal components $(X_1, X_2, \ldots, X_p)$ not amenable to perfect control. Using the notation of Pearl (2000), there is no clear, non-trivial, physical method for enacting $do(x)$, i.e. setting variable(s) X to a particular value(s) x. Such cases are common in the natural and social sciences, to say nothing of text data and spatial processes captured at a coarse resolution. To continue with the medical example, researchers often design a therapy to target one or several hub genes in full awareness that this may spur unintended interactions with other biological processes. In such a study, researchers want to learn not just whether the drug is effective but how variability in patient response can be explained by elements of X (perhaps combined with effect modifiers of pre-treatment variables Z).

This gives operational meaning to the notion of X as a *cause*

[1]Department of Computer Science, University of Oxford, Oxford, UK [2]The Alan Turing Institute, London, UK [3]Department of Statistical Science, University of College London, London, UK [4]Department of Computer Science, University of College London, London, UK. Correspondence to: Limor Gultchin <limor.gultchin@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of Y: even if $do(x)$ is undefined, we are interested in framing the effect on Y by some treatment W where the design of W comes from postulated or conjectured mechanisms triggered by X. Under the assumption that X fully *mediates* the actions W (in a technical sense defined in Sect. 2), an invariant relationship between X and Y under W becomes a useful building block for predicting the outcomes of new interventions. Furthermore, understanding *which* components of X simultaneously covary with W and Y is of independent interest, as this may suggest new targeted interventions that operate on those elements of X.

In this work, we discuss a notion of *pragmatic mediation* that provides a solution to the problem of prediction from new interventions. Our main contributions are threefold: (1) We formalize a general problem in which complex object X causes outcomes Y as a result of crude interventions W, with applications to domains with structured, high-dimensional data. (2) We describe an efficient method for estimating responses to new interventions, tractably marginalizing over the complex object X. (3) We propose a methodology for identifying practical causal mediation paths, which can provide insight into complex systems and suggest new hypotheses for future experiments.

2. Problem Setup

Let Y be an outcome of interest, and let X be postulated *causes* of Y, in the sense that hypothetical interventions on X would alter the distribution of Y (Woodward, 2003). In the machine learning and artificial intelligence literature, this is typically operationalized in terms of the interventional distribution $p(y \mid do(x))$ (Pearl, 2000). In many domains, however, perfect control is ambiguous or unattainable (VanderWeele and Hernán, 2013). This is often the case when X is a composition of more fundamental variables, as in the examples discussed in Sect. 1, as well in image and text data. We will explore the latter two in our experiments in Sect. 4.

Actionable variables and their use. In our setup, actions that change the distribution of X are assumed to exist. We index them by W, allowing this to be a random vector. Pretreatment variables Z, which are realized before $\{W, X, Y\}$, are also allowed. We say that W are our *actionable* variables, in the sense that in principle we can set them to exact values by an intervention.

By way of contrast, in the instrumental variable scenario, the target is $p(y \mid do(x))$, with W acting as an instrument that is not fundamental to the estimand of interest. If $do(x)$ is not defined and we are primarily driven by policy questions (e.g., choosing an optimal value for W), then W arguably makes the notion of X as a cause redundant. For example, Gelman (2009) suggests interpreting X as little more than a qualifier for our actions W. This is not satisfactory for the many applications where W was chosen because we expect that changing X will also change Y, even if this notion of propagation is unclear when $do(x)$ is undefined. In particular, there are practical scenarios where assumptions about invariances involving X aid the learning and prediction of policy outcomes. That is our motivation here.

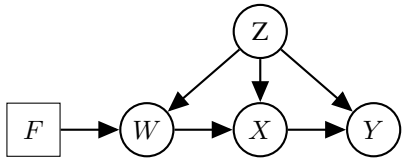

Figure 2. A DAG encoding independence assumptions in the set $\{F, W, X, Y, Z\}$, where random variables are circular vertices and intervention variable F is a square vertex. This diagram captures conditional independencies assumed in our setup, but is not Pearlian, as $do(x)$ is undefined. We cover all members of the Markov equivalence class of this graph, including those with unmeasured confounding between X and Z. We do not consider other unobserved confounders, though we discuss this in Sect. 5 as a direction for future work.

Structural assumptions. Our goal is to predict Y under intervention levels $w^\star$ of W that we have not yet seen. Following Dawid (2020), we introduce a *regime indicator* F, which is not a random variable but instead indexes the conditional distributions $p_F(w \mid z)$, with values ranging over possible interventions on W. Following Pearl (2000), we use the symbol "$do(w)$" $\in \mathcal{F}$ to mean the distribution in which $W = w$ given any $Z = z$. Our goal is then to predict Y given Z under $F = do(w)$, in particular reporting $\mathbb{E}_{F=do(w^\star)}[Y \mid z]$ for some new $w^\star$. In what follows, we will use the Pearlian notation $\mathbb{E}[Y \mid do(w), z]$ to represent the interventional distribution. Note that regime indicators can also accommodate non-atomic interventions – e.g., idle or stochastic regimes (Correa and Bareinboim, 2020) – although we will not make use of this in the sequel.

As is well-understood in the causal modeling literature, assumptions of independence between interventions and random variables can be represented by a directed acyclic graph (DAG). Our structural assumptions are encoded in Fig. 2. This is *not* a causal graph in the sense of Pearl (2000), as we are not making any claims about, say, lack of hidden common causes between Z and X or the applicability of do interventions on all variables. Instead, the DAG represents conditional independence claims such as $F \perp\!\!\!\perp Y \mid \{W, Z\}$, a statement interpretable as the lack of unmeasured confounding between W and Y given Z.[1]

Of particular importance to what follows is the implication

[1] A similar device is used, for instance, in the proof of the back-door adjustment, Thm. 3.3.2 of Pearl (2000); and earlier graphical notions of unconfoundedness, e.g. Fig. 3.19 of Spirtes et al. (1993).

$\{F, W\} \perp\!\!\!\perp Y \mid \{X, Z\}$. This conditional independence relationship, visually apparent from the d-separation in Fig. 2, informs us that there is no direct effect of W on Y. This invariance is the key point that pools together data collected at different values of F.

Problem statement: learning with pragmatic mediation. *Let $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$. Moreover, let $\mathcal{D}_l$ denote m' "labeled" datasets*

$$\mathcal{D}_l \equiv \{(W, X, Y, Z)^{l_1}, \ldots, (W, X, Y, Z)^{l_{m'}}\},$$

where $\mathcal{L} \equiv (l_1, \ldots, l_{m'}) \subset [m]$, and let $\mathcal{D}_u$ denote m'' "unlabeled" datasets

$$\mathcal{D}_u \equiv \{(W, X, Z)^{u_1}, \ldots, (W, X, Z)^{u_{m''}}\},$$

where $\mathcal{U} \equiv (u_1, \ldots, u_{m''}) \subset [m]$. In particular, $(W, X, Y, Z)^{l_i}$, $l_i \in \mathcal{L}$, denotes data collected under regime f_{l_i}, the analogous holding for $u_i \in \mathcal{U}$. Given $(\mathcal{D}_l, \mathcal{D}_u)$, the goal is to return an estimate of

$$f(w^\star, z) \equiv \mathbb{E}[Y \mid do(w^\star), z]. \tag{1}$$

The problem concerns evaluating outcomes under an intervention that sets W to $w^\star$. The *post-treatment* variables X cannot be used as a basis for decision-making, but the invariances encoded by the lack of edges $\{F, W\} \to Y$ and $F \to X$ allow for predictions of policy outcomes even in the case where pairs $(w^\star, y)$ are not in our data. As is typical of causal inference problems, we require some assumptions regarding support of treatment values in the given data. In particular, we have the following:

Assumptions (identification and support). *For all z in the support of $p(z)$: (i) the distribution $p(x \mid w^\star, z)$ is identifiable from the distributions sampled by $\mathcal{L} \cup \mathcal{U}$; (ii) the support of $p(x \mid w^\star, z)$ is contained in the union of the support of X in each dataset in $\mathcal{L}$.*

Thus, in order to obtain $\mathbb{E}[Y \mid do(w^\star), z]$ from

$$\int \mathbb{E}[Y \mid x, z] p(x \mid w^\star, z)\, dx,$$

we must have some means of generalizing to $p(x \mid w^\star, z)$ from past data, including unlabeled data. Condition (ii) says we can learn the $\mathbb{E}[Y \mid x, z]$ factor across the support of $p(x \mid w^\star, z)$ using the datasets contained in $\mathcal{L}$. This assumption can be relaxed, provided we have some principled way to extrapolate beyond the regions of (X, Z) covered by $\mathcal{L}$.

In our experiments (see Sect. 4), we predict causal responses for new interventions with no observed outcomes but some (unlabeled) data on X. Such cases arise when, for instance, Y takes a long time to be observed, or past interventions $w^\star$ took place targeting a different outcome variable. We will not constrain the functional relationship between W and X in any way, meaning that $p(x \mid w^i, z)$ need not contain any information about $p(x \mid w^j, z)$ for $i \neq j$.

This machinery operationalizes what we mean by X being a cause of Y, even if we do not define $do(x)$. At the heart of causal inference is the notion of invariance under intervention, which can be exploited even if the putative causes of interest cannot be directly manipulated. We call X a *pragmatic mediator*, i.e. a set of variables that allows us to decompose a causal model for W in Y by a model (given Z) relating W and X only, and X and Y only, under a space of interventions $\mathcal{F}$. This bears little relation to counterfactual mediation VanderWeele (2015) and demonstrates how restricted notions of mediation can be more useful than counterfactual ones in some contexts.

Related work. Although invariance principles have long been cited in formal definitions of causality Spirtes et al. (1993); Pearl (2000); Bühlmann (2020), they have recently found a new life in machine learning approaches that target more focused questions of practical interest.

The Invariant Causal Prediction method of Peters et al. (2016) – later extended by Heinze-Deml et al. (2018) and Gamella and Heinze-Deml (2020) – exploits the assumption that F does not directly cause outcome Y except for (unknown) causal parents from a candidate pool of observable variables X. There the objective is to discover the causal parents as opposed to learning what happens when we marginalize over them. Likewise, Invariant Risk Minimization (Arjovsky et al., 2020) exploits variability in F to better learn the relationship between X and Y in a way that is robust to estimation errors due to spurious, unstable associations. Again, the main focus is on the use of X, here in a (non-causal) prediction problem.

Post-treatment variables are used to improve bandit optimization by factoring the arm space, where variables that are themselves targets of interventions exhibit some (at least partially known) structure Lattimore et al. (2016); Lee and Bareinboim (2018); de Kroon et al. (2020). In contrast, we focus on the class of problems where there is little to be gained by exploiting the inner structure of X, as in many applied scenarios they are composite variables with ambiguous fine causal structure Arnold et al. (2020).

Domain adaptation, in particular covariate shift, has a long tradition of being analyzed in the context of causal models (e.g., Zhang et al. (2013)). The emphasis of this literature is how to better cope with changes in distribution between, say, training and test regimes. Although techniques such as sample reweighting for improved statistical efficiency are relevant when dealing with multiple regimes, this will not be our focus here. Instead, in Sect. 3, we emphasize convenient parametrizations of our causal setup so as to facilitate marginalization over X and parameter learning.

Finally, this work is particularly influenced by the literature on ambiguous or undefined interventions Spirtes and Scheines (2004); VanderWeele and Hernán (2013); Lee and Bareinboim (2019) as well as causal abstractions and compositional data Chalupka et al. (2017); Beckers and Halpern (2019); Arnold et al. (2020). The idea of pragmatic mediators is essentially the grounding of a causal abstraction through imperfect interventions W and the delimitation of its possible values as allowed by a "causal dictionary" $\mathcal{F}$. The ambiguity of X requires explicit assumptions about the space of modifications that we expect to enact on X.

3. Method

Based on the assumptions introduced above, this section describes (a) an algorithm for learning expected outcomes Y under crude interventions on X, operationalized as $F = do(w)$, conditioned on pre-treatment covariates Z; and (b) a procedure for interpreting the elements of X that play a mediating role in the fitted model.

Causal response estimation. We assume access to a set of features $\phi_i : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. These features are *candidate mediators*, moderated by covariates Z, which describe the outcome model for Y as

$$Y = \theta_0 + \sum_{i=1}^{d} \theta_i \phi_i(X, Z) + \epsilon, \tag{2}$$

where ϵ is an independent error term with $\mathbb{E}[\epsilon] = 0$.

Candidates may come from domain experts (e.g., experimentally validated regulatory pathways) or a data-driven approach (e.g., latent factors learned by an autoencoder). They represent a macro-level summary that clarifies what the existing $\mathcal{F}$ is able to modify in X that simultaneously contributes to Y. For example, if X describes a spatially-distributed object, like neural activations or environmental sensors, features ϕ_i can correspond to smoothing windows with localized information. If X is a text document, ϕ_i may represent aggregate interpretable interactions of relevant entities, topics, and other parts of speech. The linear assumption is substantive but not especially restrictive, given a sufficiently flexible library of basis functions Φ, which, as mentioned above, can be trained directly via neural networks or some other representation learning method.[2]

Given Eq. (2), it follows by the assumptions encoded in Fig. 2 and by linearity of expectation that

$$\mathbb{E}[Y \mid do(w), z] = \theta_0 + \sum_{i=1}^{d} \theta_i \mathbb{E}[\phi_i(X, Z) \mid w, z]. \tag{3}$$

We therefore propose a two-stage procedure to estimate $\mathbb{E}[Y \mid do(w), z]$:

1. Learn $g_i(w, z) \equiv \mathbb{E}[\phi_i(X, Z) \mid w, z]$ for all i via any black-box regression algorithm, and let $\hat{\mathbf{g}}$ denote the d-dimensional vector of resulting expectations.

2. Learn $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[(Y - \boldsymbol{\theta}^\top \hat{\mathbf{g}})^2]$ via regularized regression (e.g. Lasso, Tibshirani, 1996), to provide sparsity on $\boldsymbol{\theta}$ where supported by the data.

The procedure is detailed in Alg. 1, where we consider the case in which labeled datasets are pooled together into a set with n samples, and we learn a model for $p(x \mid w^\star, z)$ from unlabeled conditions with a single treatment level $w^\star$. This exploits the known structural relationship between W, X, Φ and Y. In particular, it represents the marginalization of X directly in terms of $\mathbb{E}[\phi_i(X, Z) \mid w, z]$,[3] which avoids the density estimation problem of learning $p(x \mid w, z)$.

There is a relation between this idea and methods for estimating non-linear causal effects in additive-error instrumental variable models based on (potentially infinite) basis expansions Singh et al. (2019); Muandet et al. (2020). However, given the potential high-dimensionality of X and the desire for interpretability, we favor dictionaries that are either hand-constructed or the result of adaptive algorithms. Moreover, although we have the option of fitting θ by regressing Y directly on Φ, we still favor the regression on $\hat{\mathbf{g}}$ instead, as $\phi_i x, z$ is a random variable not observable at test time.

Explainable pragmatic mediation. Under the assumptions of our setup, we would like to provide practitioners with qualitative information on the estimated role of the candidate mediators. Informally, we say that $\phi_i(X, Z)$ is a *causal pragmatic mediator* if and only if it covaries with W and Y simultaneously, with adjustments for Z and the other candidate mediators depending on the scenario. More formally, causal pragmatic mediators satisfy two criteria:

(i) $\phi_i(X, Z) \not\perp\!\!\!\perp W \mid Z$,

(ii) $\phi_i(X, Z) \not\perp\!\!\!\perp Y \mid \{\Phi_{\setminus i}, Z\}$

where $\Phi_{\setminus i} \equiv \Phi \setminus \phi_i(X, Z)$. We will henceforth refer to (i) and (ii) as $\mathcal{M}$-criteria.

Another way of interpreting this is by saying that W has a "nonzero conditional total effect" on ϕ_i for some $Z = z$ (that is, conditional association without adjusting for $\Phi_{\setminus i}$), while ϕ_i has a "direct effect" on Y (conditional association, also conditioning on $\Phi_{\setminus i}$).

[2] Note that, though each ϕ_i is a function of X and Z, we occasionally simplify notation by suppressing the dependence, writing ϕ_i for $\phi_i(X, Z)$ and $\Phi = \{\phi_i\}_{i=1}^d$.

[3] This can be even further simplified if we opt for product features of the shape $\phi_i(X, Z) \equiv \phi_{ix}(X)\phi_{iz}(Z)$, as in this case we have $\mathbb{E}[\phi_{ix}(X)\phi_{iz}(Z) \mid w, z] = \mathbb{E}[\phi_{ix}(X) \mid w, z]\phi_{iz}(z)$ (Kaddour et al., 2021).

Algorithm 1 Causal Response Prediction

Require: Historic interventions $\{w_i, \Phi(x_i, z_i), z_i, y_i\}_{i=1}^{n}$, new intervention training set $\{w^\star, \Phi(x_j, z_j), z_j\}_{j=1}^{n^\star_{train}}$, new intervention test set $\{w^{\star\prime}, z'_j\}_{j=1}^{n^\star_{test}}$

Historic Interventions
Learn $g_k(w, z) = \mathbb{E}[\phi_k(X, Z) \mid w, z]$
▷ Stage 1, via any black-box model
Learn $f(w, z) = \mathbb{E}[Y \mid \mathbf{g}(w, z)]$
▷ Stage 2, via an L_1-penalized model

New Intervention
Learn on train split
$g^\star_k(w^\star, z) = \mathbb{E}[\phi_k(X, Z) \mid w^\star, z]$
▷ Stage 1, update for new intervention $w^\star$
Predict on test split
$\hat{y} = f(\mathbf{g}^\star(w^{\star\prime}, z')) = \mathbb{E}[Y \mid do(w^{\star\prime}), z']$
▷ Stage 2, predict using pre-learned f
return $\hat{y}$

Algorithm 2 Pragmatic Mediation Selection

Require: Weights θ, training set $\{w_i, \Phi(x_i, z_i), z_i\}_{i=1}^{n}$, test set $\{w'_i, \Phi'(x'_i, z'_i), z'_i\}_{i=1}^{n'}$, one-sided paired difference test $c(\cdot)$, level α, mediators $\mathcal{M} = \{\}$

for $\phi_i \in \Phi$ **do**
  if $\theta_i \neq 0$ **then**
    Learn $g_i^0(z) = \mathbb{E}[\phi_i(X, Z) \mid z]$ on train split
    Learn $g_i^1(z, w) = \mathbb{E}[\phi_i(X, Z) \mid z, w]$ on train split
    Obtain residual $\epsilon_i^0 = \phi'_i - g_i^0(z')$ on test split
    Obtain residual $\epsilon_i^1 = \phi'_i - g_i^1(z', w')$ on test split
    $p = c(|\epsilon_i^0|, |\epsilon_i^1|)$
    if $p \leq \alpha$ **then**
      Add mediator $\mathcal{M} = \mathcal{M} \cup \{\phi_i\}$
    end if
  end if
end for
return $\mathcal{M}$

This definition is entirely agnostic to any possible causal structure among the elements of Φ, a structure which is itself indeterminate since $do(x)$ is not defined. Notice that the idea of combining a "total" effect "into" Φ with a "direct" effect "out of" Φ relates to settings where we may want to design new elements of $\mathcal{F}$ that "short-circuit" the mechanism, by directly targeting ϕ_i if this is at all possible and desirable in a particular domain.[4]

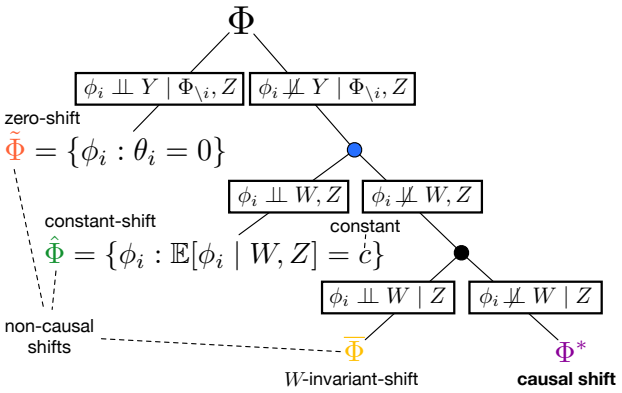

Figure 3. Recursive partition of Φ by how elements do or do not shift conditional average treatment effects. See text for details.

Although the distinction is not crucial for prediction, causal mediators can provide valuable insights about what in X characterizes the effect of W on Y. For instance, if only a subset of regions of the brain respond to stimuli and predict some behavior, then novel interventions can be designed targeting just those regions with detectable causal impact.

The leaf nodes of the tree depicted in Fig. 3 correspond to candidate mediators with different functional roles.

1. $\tilde{\Phi} \equiv \{\phi_i : \phi_i \perp\!\!\!\perp Y \mid \Phi_{\setminus i}, Z\}$. These candidates will receive zero weight in the linear formula described by Eq. (2) (and, hence, also Eq. (3)). That is, for each $\phi_i \in \tilde{\Phi}, \theta_i = 0$.

2. $\hat{\Phi} \equiv \{\phi_i : \phi_i \notin \tilde{\Phi} \wedge \phi_i \perp\!\!\!\perp W, Z\}$. In this subset, $\mathbb{E}[\phi_i \mid w, z]$ is constant for all w and z. These terms will be absorbed into the intercept of the linear formula described by Eq. (3). That is, $\mathbb{E}[Y \mid do(w), z] = \theta_0 + \sum_{\phi_i \in \hat{\Phi}} \mathbb{E}[\phi_i]$ + "function of w and z".

3. $\overline{\Phi} \equiv \{\phi_i : \phi_i \notin \{\tilde{\Phi} \cup \hat{\Phi}\} \wedge \phi_i \perp\!\!\!\perp W \mid Z\}$. These candidates will receive nonzero weight in Eq. (2), but only through the $Z \to \phi_i \to Y$ path. They are invariant in W and therefore, just like $\tilde{\Phi}$ and $\hat{\Phi}$, do not contribute to conditional average treatment effects $\mathbb{E}[Y \mid do(w), z] - \mathbb{E}[Y \mid do(w'), z]$.

4. $\Phi^* \equiv \{\phi_i : \phi_i \in \Phi \backslash \{\tilde{\Phi} \cup \hat{\Phi} \cup \overline{\Phi}\}\}$. Only this latter subclass satisfies the $\mathcal{M}$-criteria, picking out causal mediators ϕ_i on the $W \to \phi_i \to Y$ path.

This recursive partitioning of Φ immediately suggests a practical method for pragmatic mediation discovery. First, we perform our two-step estimation procedure. Then, for each ϕ_i such that $\theta_i \neq 0$, perform a conditional independence

[4]For instance, ignoring Z for simplicity: if there exists a "Pearlian" causal chain $X_1 \to X_2 \to X_3 \to Y$ in the system with no further edges, and we have a rich dictionary Φ, (features of) neither X_1 nor X_2 alone would qualify as causal pragmatic mediators, while (features of) X_3 would, even if all interventions in $\mathcal{F}$ can only directly modify X_1 and X_2.

test against the null hypothesis $H_0 : \phi_i \perp\!\!\!\perp W \mid Z$.[5] See Alg. 2 for details.

There exists no uniformly valid conditional independence test for continuous conditioning variables (Shah and Peters, 2020). However, numerous nonparametric methods have been developed with good performance on real and synthetic datasets (Heinze-Deml et al., 2018). In our experiments, we use a simple nested regression procedure, in which we compare the absolute value of out-of-sample residuals for null and alternative models – i.e., $g_i^0(z) = \mathbb{E}[\phi_i \mid z]$ and $g_i^1(z, w) = \mathbb{E}[\phi_i \mid z, w]$, respectively – using a one-sided Wilcoxon rank-sum test.[6] If predictive accuracy significantly improves with the inclusion of W, then we reject H_0. Estimation and testing are performed on separate samples to ensure unbiased inference. The procedure can easily be modified to adjust for multiple testing.

4. Experiments

In this section, we demonstrate our method in a variety of domains. We start with a simulated visual simulation task to provide a more concrete intuition for our approach. We then introduce a text data example where users are asked to edit news headlines to make them more humorous. Finally, we describe a genomics experiment where we simulate the effects of gene knockouts on the *E. coli* transcriptome. The code to reproduce results can be found at `https://github.com/limorigu/ComplexCauses`.

(a) (b) (c)

Figure 4. Visual example for image perturbation dataset. (a) Example image (z). (b) Same image, post-perturbation (x, in response to w). (c) Same image under a new perturbation regime, which we leave for the test set (x', in response to w').

4.1. Setup

We have two primary goals: (a) causal response prediction, and (b) identification of causal pragmatic mediators. We describe the overall setup for all domains below.

[5] In randomized trials, where $Z \perp\!\!\!\perp W$ by design, this can be replaced by a marginal association test against $H_0 : \phi_i \perp\!\!\!\perp W$, for those ϕ_i which are non-trivial functions of X.

[6] Other tests could in principle be substituted here, e.g. the binomial test or z-test, depending on what assumptions one is willing to make about residual distributions. See (Lei et al., 2018, Sect. 6).

Prediction. We assume access to $m-1$ mutually independent historic training regimes with corresponding labeled datasets $\mathcal{D}_{l_1}, \ldots, \mathcal{D}_{l_{m-1}}$, where each $\mathcal{D}_{l_k} = \{(W_i, X_i, Y_i, Z_i)^{l_k}\}_{i=1}^{|\mathcal{D}_{l_k}|}$. Our goal is to learn $\mathbb{E}[Y \mid do(w^\star), z]$ for a new regime $F = do(w^\star)$ (e.g., a prospective intervention). In this regime, we are given access to limited labeled training data $\mathcal{D}_{l_{w^\star}} = \{(W_i^\star, X_i, Y_i, Z_i)^{l_{w^\star}}\}_{i=1}^{|\mathcal{D}_{l_{w^\star}}|}$ and more unlabeled training data $\mathcal{D}_{u_{w^\star}} = \{(W_i^\star, X_i, Z_i)^{u_{w^\star}}\}_{i=1}^{|\mathcal{D}_{u_{w^\star}}|}$, where $|\mathcal{D}_{u_{w^\star}}| \gg |\mathcal{D}_{l_{w^\star}}|$. This captures settings where measurements for Y are expensive, delayed, or simply unrecorded. All methods are evaluated on a test dataset $\mathcal{T}_{w^\star} = \{(W_i^\star, Y_i, Z_i)^{t_{w^\star}}\}_{i=1}^{|\mathcal{T}_{w^\star}|}$.

Baseline methods that estimate $\mathbb{E}[Y \mid do(w^\star), z]$ can only make use of the labeled dataset $\mathcal{D}_{l_{w^\star}}$, as all regimes are mutually independent. However, by exploiting structural information $\Phi(X, Z)$, we are able to leverage the invariant $p(y \mid x, z)$ distribution from prior regimes. That is, we estimate $\mathbf{g}$ and $\boldsymbol{\theta}$ from $\mathcal{D}_{l_1}, \ldots, \mathcal{D}_{l_{m-1}}$ and predict effects in new regimes using only Z and W, so our method effectively treats $\mathcal{D}_{l_{w^\star}} \cup \mathcal{D}_{u_{w^\star}}$ as a single test set.

We will compare our approach to multiple regression baselines that estimate $\mathbb{E}[Y \mid do(w^\star), z]$ as the proportion of labeled data for the new regime $\mathcal{L}_{w^\star}$ grows from 10%-100%. Specifically, we consider models from four function classes: lasso regression (linear), support vector regression (SVR), random forest (RF), and gradient boosting (GB). Default hyperparameters are used throughout; see Appendix for details. We also note that other methods that seem to share similarity with our goal, such as co-training (Blum and Mitchell, 1998) and domain adaptation (Chen et al., 2011), would not be relevant baselines for comparison as they differ significantly from our work in two ways. (1) There is a two-stage functional decoupling arising from Eq. 3 alongside variable decoupling that we aim to exploit by learning g and then f; co-training does not involve such functional decoupling. (2) We can only leverage the first stage of the decoupling in a new domain. We are not aware of any domain adaptation method that accommodates this specific notion of adaptation.

As an additional check on our performance, we further consider 100 different settings of the target Y, by sampling 100 different parameters (i.e., weights $\boldsymbol{\theta}$) for its structural equations in all three tasks (for the image perturbation example we sample 1500 settings). By demonstrating consistent results across these trials, we illustrate that our method is robust to different configurations of the target variable.

Explanation. Our method is also able to find pragmatic mediators in the complex object X. By studying the high-level descriptions ϕ_i that (a) receive nonzero weight $\theta_i \neq 0$ in the sparse regression, and (b) reject $H_0 : \phi_i \perp\!\!\!\perp W | Z$ at

Table 1. General description of experimental setups.

	ImagePert	Humicroedit	DREAM5
Z	pre-perturbation image	original news headline (GloVe avg. vector)	baseline gene expression
W	location of normal distribution for perturbation	new entity edit (GloVe vector)	transcription factor out-degree
X	post-perturbation image	edited headline (GloVe avg. vector)	post-intervention gene expression
Φ	convolution windows over X	funniness hypotheses (Hossain et al., 2019)	change in kernel eigengene
Y	intensity of pixels, linear combination of Φ	funniness score, via linear combination of Φ	linear combination of Φ

$$
\begin{aligned}
t &\sim \text{Multinomial}(\boldsymbol{p}) \\
Z &= \text{pattern}_t \\
W &\sim \text{Multinomial}(\Delta_t) \\
f_w &= \begin{cases} \text{for } i = 0 \text{ to } 1000: \\ \quad \gamma \sim \ \mathcal{N}(W, \mathbf{I}) \\ \quad \text{if } (d_0, d_0) < \gamma < (d_n, d_n): \\ \quad\quad f_w[\gamma] = f_w[\gamma] + \eta \end{cases} \\
X &= Z + f_w + \mathcal{N}(0, 0.5) \\
\Phi &= \text{Convolution}_t(X) \\
Y &= \boldsymbol{\theta}^\top \Phi + \mathcal{N}(0, 0.1)
\end{aligned}
$$

Figure 5. Description of the generative model used in the experiment of Sect. 4.2.

some prespecified level α, we can identify causal mediators of relevance. We report mediator discovery error rates for all experiments below. Significance levels for all tests were fixed at $\alpha = 0.01$, with p-values adjusted for multiple testing via Holm (1979)'s method.

4.2. Image Perturbation Simulation

Setup description. Our first example is simulated and visual, which we hope will provide some intuition for the structure of this problem. We start with five possible pixel patterns (Z) and perform interventions by adding bivariate normal noise with location W. These treatments are influenced with some probability by Z. The resulting post-perturbation image (X) is then summarized via four different convolution windows, $\Phi(X, Z) = \{\phi_1, \phi_2, \phi_3, \phi_4\}$, where each ϕ_i corresponds to a quadrant of the image, and the convolution weights are indexed by the pattern corresponding to the original image Z. Finally, the intensity of the pixels over the whole image leads to an outcome (Y), given by a linear combination of Φ. The generative model used to produce the simulation is described in Fig. 5.

$\boldsymbol{p} = [0.2, 0.2, 0.2, 0.2, 0.2]$ defines the multinomial distribution from which we sample shape indicator t. Δ denotes a 5×4 matrix, where each row is a simplex containing different probabilities for selecting W values. $d_0 = 0$ and $d_n = 10$ define the dimensions of all images. The condition involving them and γ checks whether the sampled location falls within the image size. η is a perturbation parameter (fixed at 0.1 in our experiment) that is added to the sampled location if it passes the check above. This example is designed for demonstrative purposes, and ϕ_1 is constructed to be the pragmatic mediator we intend to find, as it both varies with W and has a nonzero coefficient ($\theta_1 = 0.7$) in the structural equation for Y. See full details in the Appendix. Fig. 4 shows an example set of sampled images.

Results. The results of all methods on a new intervention $w^\star$ are presented in Fig. 6. Additionally, Fig. 7 shows the process of mediator explanation for our method. The true mediator in this simulation, ϕ_1, is indicated in black. Conditional independence tests identify three windows – $\Phi = \{\phi_1, \phi_2, \phi_3\}$, indicated in red – that vary with W after conditioning on Z. We fit a lasso regression to estimate causal effects (see Eq. 3), selecting windows $\Phi = \{\phi_1, \phi_4\}$, indicated in blue. Finally, the intersection of these two sets, $\Phi^* = \{\phi_1\}$, is our causal mediator, marked in purple. Fig. 8 presents performance over 1500 different samples of parameters in the structural equation of prediction target Y. It shows similar trends to the single Y setting, where our method dominates performance by baselines until 30-40% of labels are available. The mean squared error (MSE) in this simplified example is far smaller, and required more samples to make std. error scale accordingly. We further note that for the single Y setting, we picked $\theta = \{0.7, 0, 0, -0.5\}$ to clearly demonstrate the idea of pragmatic mediation. For Fig. 8, we instead sampled $\boldsymbol{\theta}$ from a distribution (See Appendix for details), which seemed to help the performance of some baselines, while adversely affecting others.

4.3. Humorous Edits to News Headlines

Setup description. As a second example, we consider a dataset from a computational humor experiment. Participants were given news headlines and asked to make single entity changes such that the resulting headline would be humorous (Hossain et al., 2019). This work was further extended into a SemEval2020 task, and full datasets were made publicly available.[7]

[7] See `https://www.cs.rochester.edu/u/nhossain/humicroedit.html`.

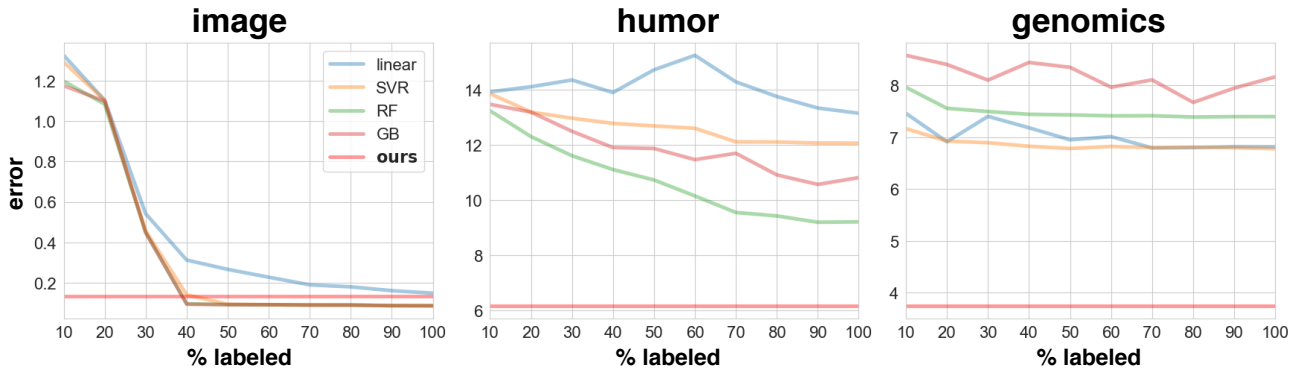

Figure 6. The mean squared error (MSE) between the estimated causal effect and the true causal effect as a function of the amount of labeled data that is available in the new regime $do(w^\star)$.

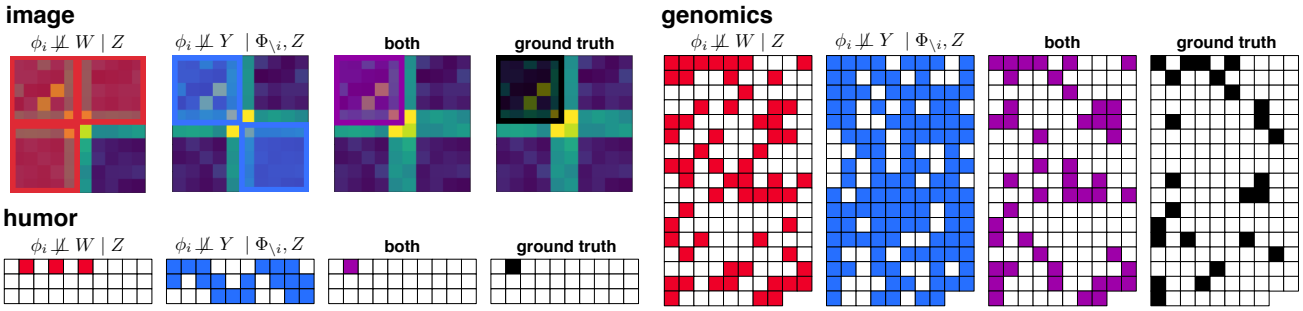

Figure 7. The ϕ_i selected by the mediation discovery method. Each set is identified as follows: red by a conditional independence test, blue by sparse regression, and purple those satisfying both tests (**black** are the true mediators). Note for the high-dimensional genomics dataset the ϕ_i are identified by only testing those ϕ_i selected by sparse regression to increase testing power.

For our evaluation, we combine all listed datasets and define the following: original headline (Z), new word introduced by edit (W), revised headline (X). Following the analysis of Hossain et al. (2019), we carried out the following pre-processing procedures: 1. We generated clusters of edit words (granular W) by performing a k-means clustering on GloVe vector representations (Pennington et al., 2014) of each edit word, with $k = 20$. The aim was to reduce the space of possible interventions to topics rather than individual words, for the purpose of defining data subsections as historic and new intervention splits. We used the resulting cluster label to create these splits. 2. We created 30 high-level descriptions ϕ for this setting (full description in the Appendix). One can think of Φ in this scenario as hypotheses to explain the funniness of an edited headline (X). 3. Computational humor is known to be a difficult domain for direct prediction tasks. For the illustrative purpose of this paper, we generated funniness scores for the outcome variable Y as a linear combination of Φ with additive noise. A random third of the coefficients are assigned a value of 0, with the rest sampled from a uniform distribution $\mathcal{U}(-1, 1)$.

Results. The results of our estimation method of the outcome Y for a random new intervention $w^\star$ are presented in Fig. 6. As can be seen, we achieve a MSE of 5.33, well below alternative estimation methods of $\mathbb{E}[Y \mid do(w^\star), z]$. Furthermore, our method correctly identified the mediator ϕ_2 in this setting, see Fig. 7. Fig. 8 provides another angle on the quality of predictions with our method by examining results over 100 trials with different configurations of $\boldsymbol{\theta}$. We can clearly see that our method still outperforms the baseline alternatives, and sees little variation in performance across parameter values, as can be seen by the small std. error bars.

4.4. Gene Knockouts

Setup description. As a final experiment, we consider semi-simulated gene knockouts based on data from the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge (Marbach et al., 2012). The *E. coli* transcriptome published as part of the DREAM5 challenge comprises a 805×4511 gene expression matrix, with 334 candidate transcription factors.[8] We use GENIE3 (Huynh-Thu et al., 2010), a leading gene regulatory network inference algorithm based on random forests, to fit the 4177 structural equations that govern this system. We treat the resulting model as our ground truth SCM.

We simulate $n = 10^4$ samples of baseline expression data for the transcription factors from a multivariate Gaussian

[8] See `http://dreamchallenges.org/project/dream-5-network-inference-challenge/`.

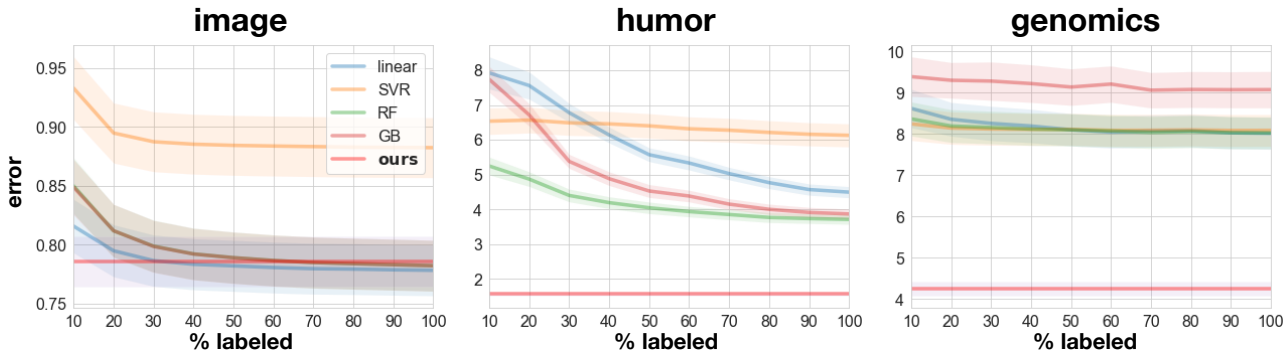

Figure 8. The mean squared error (MSE) between the estimated causal effect and the true causal effect as a function of the amount of labeled data that is available in the new regime $do(w^\star)$. Means and std. error are over 1500 for the Image Perturbation experiments, and 100 for the other two, different configurations of $\boldsymbol{\theta}$, the parameters in the structural equations giving rise to Y.

distribution with parameters estimated via maximum likelihood. These values are then propagated by GENIE3 to downstream variables, resulting in a complete set of baseline expression data (Z). We simulate 10 gene knockout experiments, summarized by the out-degree of the corresponding transcription factor (W). Post-intervention expression is once again simulated by GENIE3 (X). We treat each subnetwork of at least 10 genes as a pathway, and summarize its expression by taking the first kernel principal component (Schölkopf et al., 1999) of the corresponding submatrix, i.e. the kernel eigengene. The difference between post- and pre-intervention eigengene expression for all 168 modules that meet this dimensionality criterion constitutes our high-level summary (Φ). Modules are subsequently ranked by their Spearman correlation with W. The top and bottom 25 are assigned nonzero weight in a linear simulation of outcomes Y, with standard normal noise. More details can be found in the Appendix.

Results. Results for a random new gene knockout are presented in Fig. 6. The sparsity of this problem poses a particular challenge for baseline regression methods, which could potentially be mitigated with further tuning. In addition to achieving low MSE on the test set, our method additionally recovers 92% of all true mediators, with an overall accuracy rate of 85%. Most of the errors in this example appear to derive from false positives in the lasso regression, which could likely be improved with more cautious tuning of the λ parameter that controls model sparsity. As can be seen in Fig. 8, the same trends remain in place when repeating the experiment over 100 different configurations of $\boldsymbol{\theta}$.

5. Conclusion

In this work, we proposed a general problem setup with applications in various fields of scientific study and policy design. We showed its relevance to different modalities and domain subjects, and developed a general estimation framework. This enables the study of causal effects of crude interventions on high-dimensional, complex objects that impact an outcome of interest via some high-level mediator(s). Our approach is useful when one wishes to estimate the effects of new interventions for which little labeled data is available. We further showed how such a method can illuminate the underlying causal structure governing the process by identifying pragmatic mediation pathways between the complex objects and the outcome. Future work could extend this approach to various tasks, including: estimation of causal effects in response to high-dimensional and/or soft interventions; hypothesis design via search for interventions predicted to achieve the outcome of interest; and prediction and mediation analysis with latent variables or partial knowledge of the true causal graph.

Limitations. We see this work as a first step in the study of complex causes and crude interventions, which are distinct from the atomic, soft, or stochastic interventions that have been previously studied. Though our method is focused on a particular problem setup, we have argued that a wide variety of problems share a similar structure. Additional work could make the method more applicable in settings with unobserved confounders, or where known relationships within objects (e.g., spatial characteristics) could be exploited for greater sensitivity. Other interesting extensions of this work could examine cases where X does not fully mediate W, or when the set of abstract features Φ is not fully known. For the former, we believe that if the direct effect of W on Y is sufficiently weak, it should still be possible to exploit X as mediator. For the latter, we envision adding smoothness constraints or parametric assumptions on Φ, or simultaneously learning the abstract features, as in (Xu et al., 2020).

Acknowledgements

This work was partially supported by ONR grant 62909-19-1-2096 to DSW and RS.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv* preprint, 1907.02893, 2020.

Kellyn F. Arnold, Laurie Berrie, Peter W. G. Tennant, and Mark S. Gilthorpe. A causal inference perspective on the analysis of compositional data. *Int. J. Epidemiol.*, 49: 1307–1313, 2020.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci.*, 113(27): 7345–7352, 2016.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Peter Bühlmann. Invariance, Causality and Robustness. *Statist. Sci.*, 35(3):404–426, 2020.

Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised Discovery of El Niño Using Causal Feature Learning on Microlevel Climate Data. In *Proceedings of the Thirty-Second International Conference on Uncertainty in Artificial Intelligence*, pages 72–81, 2016.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Nips*, volume 24, pages 2456–2464. Citeseer, 2011.

J. Correa and E. Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

A. P. Dawid. Decision-theoretic foundations for statistical causality. *arXiv* preprint, 2004.12493, 2020.

Arnoud A. W. M. de Kroon, Danielle Belgrave, and Joris M. Mooij. Causal discovery for causal bandits utilizing separating sets. *arXiv*:2009.07916, 2020.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 2020.

A. Gelman. How to think about instrumental variables when you get confused. `https://statmodeling.stat.columbia.edu/2009/07/14/how_to_think_ab_2/`, 2009. Accessed: 02-02-2021.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant Causal Prediction for Nonlinear Models. *J. Causal Inference*, 6(2), 2018.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. Stat. Theory Appl.*, 6(2):65–70, 1979.

Nabil Hossain, John Krumm, and Michael Gamon. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 133–142, 2019.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5 (9):1–10, 2010.

Jean Kaddour, Qi Liu, Yuchen Zhu, Matt J. Kusner, and Ricardo Silva. Graph intervention networks for causal effect estimation. *arXiv* preprint, 2106.01939, 2021.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 4164–4172, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.*, 113(523): 1094–1111, 2018.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Kyle R Allison, Richard ... Bonneau, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nat. Methods*, 9(8): 796–804, 2012.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

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

Bernhard Schölkopf, Alexander J Smola, and Klaus-Robert Müller. Kernel Principal Component Analysis. In *Advances in Kernel Methods: Support Vector Learning*, pages 327–352, Cambridge, MA, 1999. MIT Press.

Rajen Shah and Jonas Peters. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *Ann. Statist.*, 48(3):1514–1538, 2020.

Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *J. Mach. Learn. Res.*, 9: 1941–1979, 2008.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, pages 4593–4605, 2019.

Peter Spirtes and Richard Scheines. Causal inference of ambiguous manipulations. *Philos. Sci.*, 71:833–845, 2004.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B Methodol.*, 58(1): 267–288, 1996.

Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

Tyler J. VanderWeele and Miguel A. Hernán. Causal inference under multiple versions of treatment. *J. Causal Inference*, 1:1–20, 2013.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression, 2020.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, pages 819–827, 2013.