Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

Sara Magliacane
IBM Research*
sara.magliacane@gmail.com

Thijs van Ommen
University of Amsterdam
thijsvanommen@gmail.com

Tom Claassen
Radboud University Nijmegen
tomc@cs.ru.nl

Stephan Bongers
University of Amsterdam
srbongers@gmail.com

Philip Versteeg
University of Amsterdam
p.j.j.p.versteeg@uva.nl

Joris M. Mooij
University of Amsterdam
j.m.mooij@uva.nl

Abstract

An important goal common to domain adaptation and causal inference is to make accurate predictions when the distributions for the source (or training) domain(s) and target (or test) domain(s) differ. In many cases, these different distributions can be modeled as different contexts of a single underlying system, in which each distribution corresponds to a different perturbation of the system, or in causal terms, an intervention. We focus on a class of such *causal* domain adaptation problems, where data for one or more source domains are given, and the task is to predict the distribution of a certain target variable from measurements of other variables in one or more target domains. We propose an approach for solving these problems that exploits causal inference and does not rely on prior knowledge of the causal graph, the type of interventions or the intervention targets. We demonstrate our approach by evaluating a possible implementation on simulated and real world data.
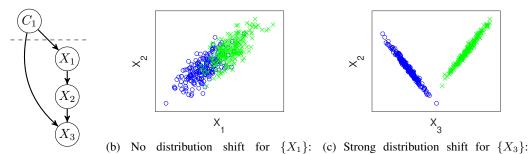
1 Introduction

Predicting unknown values based on observed data is a problem central to many sciences, and well studied in statistics and machine learning. This problem becomes significantly harder if the training and test data do not have the same distribution, for example because they come from different domains. Such a distribution shift can happen whenever the circumstances under which the training data were gathered are different from those for which the predictions are to be made. A rich literature exists on this problem of *domain adaptation*, a particular task in the field of *transfer learning*; see e.g. Quiñonero-Candela et al. [2009], Pan and Yang [2010] for overviews.

When the domain changes, so may the relations between the different variables under consideration. While for some sets of variables A, a function $f : A \to Y$ learned in one domain may continue to offer good predictions for $Y \in Y$ in a different domain, this may not be true of other sets A' of variables. *Causal graphs* [e.g., Pearl, 2009, Spirtes et al., 2000] allow us to reason about this in a principled way when the domains correspond to different external *interventions* on the system, or more generally, to different contexts in which a system has been measured. Knowledge of the causal graph that describes the data generating mechanism, and of which parts of the model are invariant

---

*Most of the work was performed while at the University of Amsterdam.

(a) Causal graph

(b) No distribution shift for $\{X_1\}$: $\mathbb{P}(Y \mid X_1, C_1 = 0) = \mathbb{P}(Y \mid X_1, C_1 = 1)$

(c) Strong distribution shift for $\{X_3\}$: $\mathbb{P}(Y \mid X_3, C_1 = 0) \neq \mathbb{P}(Y \mid X_3, C_1 = 1)$

Figure 1: In this scenario, an intervention C_1 leads to a shift of distribution between source domain and target domain (see also Example 1). Green crosses show source domain data ($C_1 = 0$), blue circles show target domain data ($C_1 = 1$). A standard feature selection method that does not take into account the causal structure, but would use X_3 to predict $Y := X_2$ (because X_3 is a good predictor of Y in the source domain), would obtain extremely biased predictions in the target domain. Using X_1 instead yields less accurate predictions in the source domain, but much more accurate ones in the target domain.

across the different domains, allows one to transfer knowledge from one domain to the other in order to address the problem of domain adaptation [Spirtes et al., 2000, Storkey, 2009, Schölkopf et al., 2012, Bareinboim and Pearl, 2016].

Over the last years, various methods have been proposed to exploit the causal structure of the data generating process in order to address certain domain adaptation problems, each relying on different assumptions. For example, Bareinboim and Pearl [2016] provide theory for identifiability under transfer ("transportability") assuming that the causal graph is known, that interventions are perfect, and that the intervention targets are known. Hyttinen et al. [2015] also assume perfect interventions with known targets but do not rely on complete knowledge of the causal graph, instead inferring the relevant aspects of it from the data. Rojas-Carulla et al. [2018] make the assumption that if the conditional distribution of the target given some subset of covariates is invariant across different source domains, then this conditional distribution must also be the same in the target domain. The methods proposed in [Schölkopf et al., 2012, Zhang et al., 2013, 2015, Gong et al., 2016] all address challenging settings in which conditional independences that follow from the usual Markov and faithfulness assumptions alone do not suffice to solve the problem, but additional assumptions on the data generating process have to be made.

In this work, we will make no such additional assumptions, and address the setting in which both the causal graph *and* the intervention types and targets may be (partially) unknown. Our contributions are the following. We consider a set of relatively weak assumptions that make the problem well-posed. We propose an approach to solve this class of causal domain adaptation problems that can deal with the presence of latent confounders. The main idea is to select the subset of features $\boldsymbol{A}$ that leads to the best predictions of Y in the source domains, while satisfying *invariance* (i.e., $\mathbb{P}(Y \mid \boldsymbol{A})$ is the same in the source and target domains). To test whether the invariance condition is satisfied, we apply the recently proposed Joint Causal Inference (JCI) framework [Mooij et al., 2018] to exploit the information provided by multiple domains corresponding to different interventions. The basic idea is as follows. First, a standard feature selection method is applied to source domains data to find sets of features that are predictive of a target variable, trading off bias and variance, but unaware of changes in the distribution across domains. A causal inference method then draws conclusions from all given data about the possible causal graphs, avoiding sets of features for which the predictions would not transfer to the target domains. We propose a proof-of-concept implementation of our approach building on a causal discovery algorithm by Hyttinen et al. [2014]. We evaluate the method on synthetic data and a real-world example.

2 Theory

Before giving a precise definition of the class of domain adaptation problems that we consider in this work, we begin with a motivating example.

Example 1. *We are given three variables X_1, X_2, X_3 describing different aspects of a system (for example, certain blood cell phenotypes in mice). We have observational measurements of these three variables (the source domain, designated with $C_1 = 0$), and in addition, measurements of X_1 and X_3 under an intervention (the target domain, designated with $C_1 = 1$), e.g., in which the mice have been exposed to a certain drug. The domain adaptation task is to predict the values of $Y := X_2$ in the interventional target domain (i.e., when $C_1 = 1$). Let us assume for this example that the causal graph in Figure 1a applies, i.e., we assume that X_2 is affected by X_1 and affects X_3, while C_1 affects both X_1 and X_3 (i.e., the intervention targets the variables X_1 and X_3). This causal graph implies $\mathbb{P}(Y \mid X_1, C_1 = 0) = \mathbb{P}(Y \mid X_1, C_1 = 1)$. Suppose further that the relation between X_1 and X_2 is about equally strong as the relation between X_2 and X_3, but considerably more noisy. Then a feature selection method using only available source domain data, and aiming to select the best subset of features to use for prediction of Y will prefer both $\{X_3\}$ and $\{X_1, X_3\}$ over $\{X_1\}$ (because predicting Y from X_1 leads to larger variance than predicting Y from X_3, and to a larger bias than predicting Y from both X_1 and X_3). However, under the intervention ($C_1 = 1$), $\mathbb{P}(Y \mid X_3)$ and $\mathbb{P}(Y \mid X_1, X_3)$ both change,[2] so that using those features to predict Y in the target domain could lead to extreme bias, as illustrated in Figure 1c. Because the conditional distribution of Y given X_1 is invariant across domains, as illustrated in Figure 1b, predictions of Y based only on X_1 can be safely transferred to the target domain.*

This example provides an instance of a domain adaptation problem where feature selection methods that do not take into account the causal structure would pick a set of features that does not generalize to the target domain, and may lead to arbitrarily bad predictions (even asymptotically, as the number of data points tends to infinity). On the other hand, correctly taking into account the causal structure and the possible distribution shift from source to target domain allows to upper bound the prediction error in the target domain, as we will see in Section 2.3.

2.1 Problem Setting

We now formalize the domain adaptation problems that we address in this paper. We will make use of the terminology of the recently proposed Joint Causal Inference (JCI) framework [Mooij et al., 2018].

Let us consider a system of interest described by a set of *system variables* $\{X_j\}_{j \in \mathcal{J}}$. In addition, we model the domain in which the system has been measured by *context variables* $\{C_i\}_{i \in \mathcal{I}}$ (we will use "context" as a synonym for "domain"). We will denote the tuple of all system and context variables as $\boldsymbol{V} = ((X_j)_{j \in \mathcal{J}}, (C_i)_{i \in \mathcal{I}})$. System and context variables can be discrete or continuous. As a concrete example, the system of interest could be a mouse. The system variables could be blood cell phenotypes such as the concentration of red blood cells, the concentration of white blood cells, and the mean red blood cell volume. The context variables could indicate for example whether a certain gene has been knocked out, the dosage of a certain drug administered to the mice, the age and gender of the mice, or the lab in which the measurements were done. The important underlying assumption is that context variables are *exogenous* to the system, whereas system variables are *endogenous*. The interventions are not limited to the perfect ("surgical") interventions modeled by the do-operator of Pearl [2009], but can also be other types of interventions such as mechanism changes [Tian and Pearl, 2001], soft interventions [Markowetz et al., 2005], fat-hand interventions [Eaton and Murphy, 2007], activity interventions [Mooij and Heskes, 2013], and stochastic versions of all these. Knowledge of the intervention *targets* is not necessary (but is certainly helpful). For example, administering a drug to the mice may have a direct causal effect on an unknown subset of the system variables, but we can simply model it as a binary exogenous variable (indicating whether or not the drug was administered) or a continuous exogenous variable (describing the dosage of the administered drug) without specifying in advance on which variables it has a direct effect. We can now formally state the domain adaptation task that we address in this work:

Task 1 (Domain Adaptation Task). *We are given data for a single or for multiple source domains, in each of which $C_1 = 0$, and for a single or for multiple target domains, in each of which $C_1 = 1$. Assume the source domains data is complete (i.e., no missing values), and the target domains data is complete with the exception of all values of a certain target variable $Y = X_j$. The task is to predict these missing values of the target variable Y given the available source and target domains data.*

---

[2]More precisely, we should say that $\mathbb{P}(Y \mid X_3, C_1 = 1)$ may differ from $\mathbb{P}(Y \mid X_3, C_1 = 0)$, and similarly when conditioning on $\{X_1, X_3\}$.
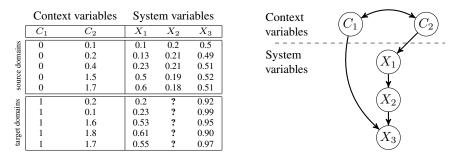
	Context variables		System variables		
	C_1	C_2	X_1	X_2	X_3
source domains	0	0.1	0.1	0.2	0.5
	0	0.2	0.13	0.21	0.49
	0	0.4	0.23	0.21	0.51
	0	1.5	0.5	0.19	0.52
	0	1.7	0.6	0.18	0.51
target domains	1	0.2	0.2	?	0.92
	1	0.1	0.23	?	0.99
	1	1.6	0.53	?	0.95
	1	1.8	0.61	?	0.90
	1	1.7	0.55	?	0.97

Figure 2: Example of a causal domain adaptation problem. The causal graph is depicted on the right, the corresponding data on the left. The task is to predict the missing values of $Y = X_2$ in the target domains ($C_1 = 1$), based on the observed data from the source domains and the target domains, without knowledge of the causal graph. See also Example 2.

An example is provided in Figure 2. In the next subsection, we will formalize our assumptions to turn this task into a well-posed problem.

2.2 Assumptions

Our first main assumption is that the data generating process (on both system and context variables) can be represented as a Structural Causal Model (SCM) (see e.g., [Pearl, 2009]):

$$
\mathcal{M} : \begin{cases} C_i & = g_i(\boldsymbol{E}_{\text{PA}(i) \cap \mathcal{K}}), & i \in \mathcal{I} \\ X_j & = f_j(\boldsymbol{X}_{\text{PA}(j) \cap \mathcal{J}}, \boldsymbol{C}_{\text{PA}(j) \cap \mathcal{I}}, \boldsymbol{E}_{\text{PA}(j) \cap \mathcal{K}}), & j \in \mathcal{J} \\ p(\boldsymbol{E}) & = \prod_{k \in \mathcal{K}} p(E_k). \end{cases} \tag{1}
$$

Here, we introduced exogenous latent independent "noise" variables $(E_k)_{k \in \mathcal{K}}$ that model latent causes of the context and system variables. The parents of each variable are denoted by $\text{PA}(\cdot)$. Each context and system variable is related to its parent variables by a structural equation. In addition, we assume a factorizing probability distribution on the exogenous variables. There could be cyclic dependencies, for example due to feedback loops, but for simplicity of exposition we will discuss only the acyclic case here, noting that the extension to the cyclic case is straightforward given recent theoretical advances on cyclic SCMs [Bongers et al., 2018]. This SCM provides a causal model for the distributions of the various domains, and in particular, it induces a joint distribution $\mathbb{P}(\boldsymbol{V})$ on the context and system variables. Note that we will assume that the data generating process can be modeled by some model of this form, but we do not rely on knowing the precise model.

The SCM $\mathcal{M}$ can be represented graphically by its causal graph $\mathcal{G}(\mathcal{M})$, a graph with nodes $\mathcal{I} \cup \mathcal{J}$ (i.e., the labels of both system and context variables), directed edges $l_1 \to l_2$ for $l_1, l_2 \in \mathcal{I} \cup \mathcal{J}$ iff $l_1 \in \text{PA}(l_2)$, and bidirected edges $l_1 \leftrightarrow l_2$ for $l_1, l_2 \in \mathcal{I} \cup \mathcal{J}$ iff there exists a $k \in \text{PA}(l_1) \cap \text{PA}(l_2) \cap \mathcal{K}$. In the acyclic case, this causal graph is an Acyclic Directed Mixed Graph (ADMG), and $\mathcal{M}$ is also known as a Semi-Markov Causal Model (see e.g., [Pearl, 2009]). The directed edges represent direct causal relationships, and the bidirected edges may represent hidden confounders (both relative to the set of variables in the ADMG). The (causal) *Markov assumption* holds [Richardson, 2003], i.e., any d-separation $\boldsymbol{A} \perp \boldsymbol{B} \,|\, \boldsymbol{S}\ [\mathcal{G}(\mathcal{M})]$ between sets of random variables $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S} \subseteq \boldsymbol{V}$ in the ADMG $\mathcal{G}(\mathcal{M})$ implies a conditional independence $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \,|\, \boldsymbol{S}\ [\mathbb{P}(\boldsymbol{V})]$ in the distribution $\mathbb{P}(\boldsymbol{V})$ induced by the SCM $\mathcal{M}$. A standard assumption in causal discovery is that the joint distribution $\mathbb{P}(\boldsymbol{V})$ is *faithful* with respect to the ADMG $\mathcal{G}(\mathcal{M})$, i.e., that there are no other conditional independences in the joint distribution than those implied by d-separation.

We will make the following assumptions on the causal structure (where henceforth we will simply write $\mathcal{G}$ instead of $\mathcal{G}(\mathcal{M})$), which are discussed in detail by Mooij et al. [2018]:

Assumption 1 (JCI Assumptions). *Let $\mathcal{G}$ be a causal graph with variables $\boldsymbol{V}$ (consisting of system variables $\{X_j\}_{j \in \mathcal{J}}$ and context variables $\{C_i\}_{i \in \mathcal{I}}$).*

 (i) *No system variable directly causes any context variable ("exogeneity")*
    ($\forall j \in \mathcal{J}, \forall i \in \mathcal{I} : X_j \to C_i \notin \mathcal{G}$);
 (ii) *No system variable is confounded with a context variable ("randomization")*
    ($\forall j \in \mathcal{J}, \forall i \in \mathcal{I} : X_j \leftrightarrow C_i \notin \mathcal{G}$);

(iii) Every pair of context variables is purely confounded ("genericity")
$$(\forall i, i' \in \mathcal{I} : C_i \leftrightarrow C_{i'} \in \mathcal{G} \wedge C_i \rightarrow C_{i'} \notin \mathcal{G}).$$

The first assumption is the most crucial one that captures what we mean by "context". The other two assumptions are less crucial and could be omitted, depending on the application. For a more in-depth discussion of these modeling assumptions and on how they compare with other possible causal modeling approaches, we refer the reader to [Mooij et al., 2018]. Any causal discovery method can in principle be used in the JCI setting, but identifiability greatly benefits from taking into account the background knowledge on the causal graph from Assumption 1.

In addition, in order to be able to address the causal domain adaptation task, we will assume:

Assumption 2. *Let $\mathcal{G}$ be a causal graph with variables $\boldsymbol{V}$ (consisting of system variables $\{X_j\}_{j \in \mathcal{J}}$ and context variables $\{C_i\}_{i \in \mathcal{I}}$), and $\mathbb{P}(\boldsymbol{V})$ be the corresponding distribution on $\boldsymbol{V}$. Let C_1 be the source/target domains indicator and $Y = X_j$ the target variable.*

(i) The distribution $\mathbb{P}(\boldsymbol{V})$ is Markov and faithful w.r.t. $\mathcal{G}$;
(ii) Any conditional independence involving Y in the source domains also holds in the target domains, i.e., if $\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{S}$ contains Y but not C_1 then:[3]

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 0] \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 1];$$

(iii) C_1 has no direct effect on Y w.r.t. $\boldsymbol{V}$, i.e., $C_1 \rightarrow Y \notin \mathcal{G}$.

The Markov and faithfulness assumptions are standard in constraint-based causal discovery on a single domain; we apply them here on the "meta-system" composed of system and context.

Assumption 2(ii) may seem non-intuitive, but as we show in the Supplementary Material, it follows from more intuitive (but stronger) assumptions, for example if both the pooled source domains distribution $\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)$ and the pooled target domains distribution $\mathbb{P}(\boldsymbol{V} \mid C_1 = 1)$ are Markov and faithful to the subgraph of $\mathcal{G}$ which excludes C_1. These stronger assumptions imply that the causal structure (i.e., presence or absence of causal relationships and confounders) of the other variables is invariant when going from source to target domains. Assumption 2(ii) is a weakened version of these more natural assumptions, allowing additional *independences* to hold in the target domains compared to the source domains, e.g., when C_1 models a perfect surgical intervention.

Assumption 2(iii) is strong, yet some assumption of that type seems necessary to make the task well-defined. Without any information at all about the target(s) of C_1, or the causal mechanism that determines the values of Y in the target domains, predicting the values of Y for the target domains seems generally impossible. Note that the assumption is more likely to be satisfied if the interventions are believed to be precisely targeted, and gets weaker the more relevant system variables are observed.[4]

As one example of a real-world setting in which these assumptions are reasonable, consider a genomics experiment, in which gene expression levels of many different genes are measured in response to knockouts of single genes. Given our present-day understanding of the biology of gene expression, it is very reasonable to assume that the knockout of gene X_i only has a direct effect on the expression level of gene X_i itself. As long as we do not ask to predict the expression level of X_i under a knockout of X_i, but only the expression level of other genes $Y = X_j$ with $j \neq i$, Assumption 2(iii) seems justified. It is also reasonable (based on present-day understanding of biology) to expect that a single gene knockout does not change the causal mechanisms in the rest of the system. This justifies Assumption 2(ii) in this setting if one is willing to assume faithfulness.

In the next subsections, we will discuss how these assumptions enable us to address the domain adaptation task.

---

[3]Here, with $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 0]$ we mean $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)]$, i.e., the conditional independence of $\boldsymbol{A}$ from $\boldsymbol{B}$ given $\boldsymbol{S}$ in the mixture of the source domains $\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)$, and similarly for the target domains.

[4]This assumption can be weakened further: in some circumstances one can infer from the data and the other assumptions that C_1 cannot have a direct effect on Y. For example: if there exists a descendant $D \in \text{DE}(Y)$, and if there exists a set $\boldsymbol{S} \subseteq \boldsymbol{V} \setminus (\{C_1, Y\} \cup \text{DE}(Y))$, such that $C_1 \perp\!\!\!\perp D \mid \boldsymbol{S}$, then C_1 is not a direct cause of Y w.r.t. $\boldsymbol{V}$. For some proposals on alternative assumptions that can be made when this assumption is violated, see e.g., [Schölkopf et al., 2012, Zhang et al., 2013, 2015, Gong et al., 2016].

2.3 Separating Sets of Features

Our approach to addressing Task 1 is based on finding a *separating set* $A \subseteq V \setminus \{C_1, Y\}$ of (context and system) variables that satisfies $C_1 \perp Y \mid A \; [\mathcal{G}]$. If such a separating set A can be found, then the distribution of Y conditional on A is *invariant* under transferring from the source domains to the target domains, i.e., $\mathbb{P}(Y \mid A, C_1 = 0) = \mathbb{P}(Y \mid A, C_1 = 1)$. As the former conditional distribution can be estimated from the source domains data, we directly obtain a prediction for the latter, which then enables us to predict the values of Y from the observed values of A in the target domains.[5]

We will now discuss the effect of the choice of A on the quality of the predictions. For simplicity of the exposition, we make use of the squared loss function and look at the asymptotic case, ignoring finite-sample issues. When predicting Y from a subset of features $A \subseteq V \setminus \{Y, C_1\}$ (that may or may not be separating), the optimal predictor is defined as the function $\hat{Y}$ mapping from the range of possible values of A to the range of possible values of Y that minimizes the *target domains risk* $\mathbb{E}\big((Y - \hat{Y}(A))^2 \mid C_1 = 1\big)$, and is given by the conditional expectation (regression function) $\hat{Y}_A^1(a) := \mathbb{E}(Y \mid A = a, C_1 = 1)$. Since Y is not observed in the target domains, we cannot directly estimate this regression function from the data.

One approach that is often used in practice is to ignore the difference in distribution between source and target domains, and use instead the predictor $\hat{Y}_A^0(a) := \mathbb{E}(Y \mid A = a, C_1 = 0)$, which minimizes the *source domains risk* $\mathbb{E}\big((Y - \hat{Y})^2 \mid C_1 = 0\big)$. This approximation introduces a bias $\hat{Y}_A^1 - \hat{Y}_A^0$ that we will refer to as the *transfer bias* (when predicting Y from A). When ignoring that source domains and target domains have different distributions, any standard machine learning method can be used to predict Y from A. As the transfer bias can become arbitrarily large (as we have seen in Example 1), the prediction accuracy of this solution strategy may be arbitrarily bad (even in the infinite-sample limit).

Instead, we propose to only predict Y from A when the set A of features satisfies the following *separating set* property:
$$C_1 \perp Y \mid A \; [\mathcal{G}], \tag{2}$$
i.e., it d-separates C_1 from Y in $\mathcal{G}$. By the Markov assumption, this implies $C_1 \perp\!\!\!\perp Y \mid A \; [\mathbb{P}(V)]$. In other words (as already mentioned above), for separating sets, the distribution of Y conditional on A is *invariant* under transferring from the source domains to the target domains, i.e., $\mathbb{P}(Y \mid A, C_1 = 0) = \mathbb{P}(Y \mid A, C_1 = 1)$. By virtue of this invariance, regression functions are identical for the source domains and target domains, i.e., $\hat{Y}_A^0 = \hat{Y}_A^1$, and hence also the source domains and target domains risks are identical when using the predictor $\hat{Y}_A^0$:
$$C_1 \perp Y \mid A \; [\mathcal{G}] \implies \mathbb{E}\big((Y - \hat{Y}_A^0)^2 \mid C_1 = 1\big) = \mathbb{E}\big((Y - \hat{Y}_A^0)^2 \mid C_1 = 0\big). \tag{3}$$
The r.h.s. can be estimated from the source domains data, and the l.h.s. equals the generalization error to the target domains when using the predictor $\hat{Y}_A^0$ trained on the source domains (which equals the predictor $\hat{Y}_A^1$ that one could obtain if all target domains data, including the values of Y, were observed).[6] Although this approach leads to zero transfer bias, it introduces another bias: by using only a subset of the features A, rather than *all available* features $V \setminus \{C_1, Y\}$, we may miss relevant information to predict Y. We refer to this bias as the *incomplete information bias*, $\hat{Y}_{V \setminus \{Y, C_1\}}^1 - \hat{Y}_A^1$.

The total bias when using $\hat{Y}_A^0$ to predict Y is the sum of the transfer bias and the incomplete information bias:
$$\underbrace{\hat{Y}_{V \setminus \{Y, C_1\}}^1 - \hat{Y}_A^0}_{\text{total bias}} = \underbrace{(\hat{Y}_A^1 - \hat{Y}_A^0)}_{\text{transfer bias}} + \underbrace{(\hat{Y}_{V \setminus \{Y, C_1\}}^1 - \hat{Y}_A^1)}_{\text{incomplete information bias}} .$$

For some problems, one may be better off by simply ignoring the transfer bias and minimizing the incomplete information bias, while for other problems, it is crucial to take the transfer into account to

---

[5]This trivial observation is not novel; see e.g. [Ch. 7, p. 164, Spirtes et al., 2000]. It also follows as a special case of [Theorem 2, Pearl and Bareinboim, 2011]. The main novelty of this work is the proposed strategy to identify such separating sets.

[6]Note that this equation only holds asymptotically; for finite samples, in addition to the transfer from source domains to target domains, we have to deal with the generalization from empirical to population distributions and from the covariate shift if $\mathbb{P}(A \mid C_1 = 1) \neq \mathbb{P}(A \mid C_1 = 0)$ [see e.g. Mansour et al., 2009].

obtain small generalization errors. In that situation, we could use any subset $\boldsymbol{A}$ for prediction that satisfies the separating set property (2), implying zero transfer bias; obviously, the best predictions are then obtained by selecting a separating subset that also minimizes the source domains risk (i.e., minimizes the incomplete information bias). We conclude that this strategy of selecting a subset $\boldsymbol{A}$ to predict Y may yield an asymptotic guarantee on the prediction error by (3), whereas simply ignoring the shift in distribution may lead to unbounded prediction error, since the transfer bias could be arbitrarily large in the worst case scenario.

2.4 Identifiability of Separating Feature Sets

For the strategy of selecting the best separating sets of features as discussed in Section 2.3, we need to find one or more sets $\boldsymbol{A} \subseteq \boldsymbol{V} \setminus \{C_1, Y\}$ that satisfy (2). Of course, the problem is that we cannot directly test this in the data, because the values of Y are missing for $C_1 = 1$. Note that also Assumption 2(ii) cannot be directly used here, because it only applies when C_1 is *not* in $\boldsymbol{A} \cup \boldsymbol{B}$. When the causal graph $\mathcal{G}$ is known, it is easy to verify whether (2) holds directly using d-separation. Here we address the more challenging setting in which the causal graph and the targets of the interventions are (partially) unknown.[7] Conceptually, one could estimate a set of possible causal graphs by using a causal discovery algorithm (for example, extending any standard method to deal with the missing conditional independence tests in $C_1 = 1$), and then read off separating sets from these graphs. In practice, it is not necessary to estimate completely these causal graphs: we only need to know enough about them to verify or falsify whether a given set of features separates C_1 from Y. The following example (with details in the Supplementary Material) illustrates a case where such reasoning allows us to identify a separating set.

Example 2. *Assume that Assumptions 1 and 2 hold for two context variables C_1, C_2 and three system variables X_1, X_2, X_3 with $Y := X_2$. If the following conditional (in)dependences all hold in the source domains:*

$$C_2 \perp\!\!\!\perp X_2 \mid X_1 \; [C_1 = 0], \qquad C_2 \not\perp\!\!\!\perp X_2 \mid \emptyset \; [C_1 = 0], \qquad C_2 \perp\!\!\!\perp X_3 \mid X_2 \; [C_1 = 0], \qquad (4)$$

then $C_1 \perp\!\!\!\perp X_2 \mid X_1 \; [\mathcal{G}]$, i.e., $\{X_1\}$ is a separating set for C_1 and X_2. One possible causal graph leading to those (in)dependences is provided in Figure 2 (the others are shown in Figure 1(a) in the Supplementary Material). For that ADMG, and given enough data, feature selection applied to the source domains data will generically select $\{X_1, X_3\}$ as the optimal set of features for predicting $Y := X_2$, which can lead to an arbitrarily large prediction error. On the other hand, the set $\{X_1\}$ is separating in any ADMG satisfying (4), so using it to predict Y leads to zero transfer bias, and therefore provides a guarantee on the target domains risk (i.e., it provides an upper bound on the optimal target domains risk, which can be estimated from the source domains data).

Rather than characterizing by hand all possible situations in which a separating set can be identified (like in Example 2), in this work we delegate the causal inference to an automatic theorem prover. Intuitively, the idea is to provide the automatic theorem prover with the conditional (in)dependences that hold in the data, in combination with an encoding of Assumptions 1 and 2 into logical rules, and ask the theorem prover whether it can prove that $C_1 \perp\!\!\!\perp Y \mid \boldsymbol{A}$ holds for a candidate set $\boldsymbol{A}$ from the assumptions and provided conditional (in)dependences. There are three possibilities: either it can prove the query (and then we can proceed to predict Y from $\boldsymbol{A}$ and get an estimate of the target domains risk), or it can disprove the query (and then we know $\boldsymbol{A}$ will generically give predictions that suffer from an arbitrarily large transfer bias), or it can do neither (in which case hopefully another subset $\boldsymbol{A}$ can be found that does provably satisfy (2)).

2.5 Algorithm

A simple (brute-force) algorithm that finds the best separating set as described in Section 2.3 is the following. By using a standard feature selection method, produce a ranked list of subsets $\boldsymbol{A} \subseteq \boldsymbol{V} \setminus \{Y, C_1\}$, ordered ascendingly with respect to the empirical source domains risks. Going through this list of subsets (starting with the one with the smallest empirical source domains risk),

---

[7]Another option, proposed by Rojas-Carulla et al. [2018], is to *assume* that if $p(Y \mid \boldsymbol{A})$ is invariant across all source domains (i.e., $p(Y \mid \boldsymbol{A}, C_1 = 0, C_{\setminus 1} = c) = p(Y \mid \boldsymbol{A}, C_1 = 0)$ for all c), then the same holds across all source *and* target domains (i.e., $p(Y \mid \boldsymbol{A}, C_1 = 1) = p(Y \mid \boldsymbol{A}, C_1 = 0, C_{\setminus 1} = c)$ for all c). This assumption can be violated in some simple cases, e.g. see Example 2.

test whether the separating set property can be inferred from the data by querying the automated theorem prover. If (2) can be shown to hold, use that subset A for prediction of Y and stop; if not, continue with the next candidate subset A in the list. If no subset satisfies (2), abstain from making a prediction.[8]

An important consequence of Assumption 2(ii) is that it enables us to transfer conditional independence involving the target variable from the source domains to the target domains (proof provided in the Supplementary Material):

Proposition 1. *Under Assumption 2,*

$$A \perp\!\!\!\perp B \mid S \, [C_1 = 0] \iff A \perp\!\!\!\perp B \mid S \cup \{C_1\} \iff A \perp\!\!\!\perp B \mid S \cup \{C_1\} \, [\mathcal{G}]$$

for subsets $A, B, S \subseteq V$ such that their union contains Y but not C_1.

To test the separating set condition (2), we use the approach proposed by Hyttinen et al. [2014], where we simply add the JCI assumptions (Assumption 1) as constraints on the optimization problem, in addition to the domain-adaptation specific assumption that $C_1 \to Y \notin \mathcal{G}$ (Assumption 2(iii)). As inputs we use all directly testable conditional independence test p-values $p_{A \perp\!\!\!\perp B \mid S}$ in the pooled data (when $Y \notin A \cup B \cup S$) and all those resulting from Proposition 1 from the source domains data only (if $Y \in A \cup B \cup S$). If background knowledge on intervention targets or the causal graph is available, it can easily be added as well. We use the method proposed by Magliacane et al. [2016] to query for the confidence of whether some statement (e.g., $Y \perp\!\!\!\perp C_1 \mid A$) is true or false. The results of Magliacane et al. [2016] show that this approach is sound under oracle inputs, and asymptotically consistent whenever the statistical conditional independence tests used are asymptotically consistent. In other words, in this way the probability of wrongly deciding whether a subset A is a separating set converges to zero as the sample size increases. We chose this approach because it is simple to implement on top of existing open source code.[9] Note that the computational cost quickly increases with the number of variables, limiting the number of variables that can be considered simultaneously.

One remaining issue is how to predict Y when an optimal separating set A has been found. As the distribution of A may shift when transferring from source domains to target domains, this means that there is a *covariate shift* to be taken into account when predicting Y. Any method (e.g., least-squares regression) could in principle be used to predict Y from a given set of covariates, but it is advisable to use a prediction method that works well under covariate shift, e.g., [Sugiyama et al., 2008].

3 Evaluation

We perform an evaluation on both synthetic data and a real-world dataset based on a causal inference challenge.[10] The latter dataset consists of hematology-related measurements from the International Mouse Phenotyping Consortium (IMPC), which collects measurements of phenotypes of mice with different single-gene knockouts.

In both evaluations we compare a standard feature selection method (which uses Random Forests) with our method that builds on top of it and selects from its output the best separating set. First, we score all possible subsets of features by their out-of-bag score using the implementation of Random Forest Regressor from `scikit-learn` [Pedregosa et al., 2011] with default parameters. For the baseline we then select the best performing subset and predict Y. Instead, for our proposed method we try to find a subset of features A that is also a separating set, starting from the subsets with the best scores. To test whether A is a separating set, we use the method described in Section 2.5, using the ASP solver `clingo 4.5.4` [Gebser et al., 2014]. We provide as inputs the independence test results from a partial correlation test with significance level $\alpha = 0.05$ and combine it with the weighting scheme from Magliacane et al. [2016]. We then use the first subset A in the ranked list of predictive sets of features found by the Random Forest method for which the confidence that $C_1 \perp\!\!\!\perp Y \mid A$ holds is positive. If there is no set A that satisfies this criterion, then we abstain from making a prediction.

---

[8]Abstaining from predictions can be advantageous when trading off recall and precision. If a prediction *has* to be made, we can fall back on some other method or simply accept the risk that the transfer bias may be large.

[9]We build on the source code provided by Magliacane et al. [2016] which in turn extends the source code provided by Hyttinen et al. [2014]. The full source code of our implementation and the experiments is available online at `https://github.com/caus-am/dom_adapt`.

[10]Part of the CRM workshop on Statistical Causal Inference and Applications to Genetics, Montreal, Canada (2016). See also `http://www.crm.umontreal.ca/2016/Genetics16/competition_e.php`

(a) Synthetic data with $N = 1000$ samples and a large perturbation
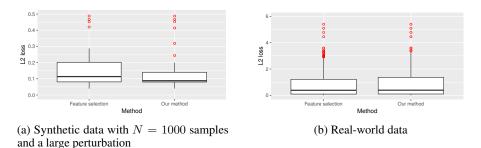
(b) Real-world data

Figure 3: Evaluation results (see main text and Supplementary Material for details).

For the synthetic data, we generate randomly 200 linear acyclic models with latent variables and Gaussian noise, each with three system variables, and sample N data points each for the observational and two experimental domains, where we simulate soft interventions on randomly selected targets, with different sizes of perturbations. We randomly select which of the two context variables will be C_1 and which of the three system variables will be Y. We disallow direct effects of C_1 on Y, and enforce that no intervention can directly affect all variables simultaneously. More details on how the data were simulated are provided in the Supplementary Material. Figure 3a shows a boxplot of the L_2 loss of the predicted Y values with respect to the true values for both the baseline and our method, considering the 121 cases out of 200 in which our method does produce an answer. In particular, Figure 3a considers the case of $N = 1000$ samples per regime and interventions that all produce a large perturbation. In the Supplementary Material we show that results improve with more samples, both for the baseline, but even more so for our method, since the quality of the conditional independence tests improves. We also show that, according to expectations, if the target distribution is very similar to the source distributions, i.e., the transfer bias is small, our method does not provide any benefit and seems to perform worse than the baseline. Conversely, the larger the intervention effect, the bigger the advantage of using our method.

For the real-world dataset, we select a subset of the variables considered in the CRM Causal Inference Challenge. Specifically, for simplicity we focus on 16 phenotypes that are not deterministically related to each other. The dataset contains measurements for 441 "wild type" mice and for about 10 "mutant" mice for each of 13 different single gene knockouts. We then generate 1000 datasets by randomly selecting subsets of 3 variables and 2 gene knockout contexts, and always include also "wild type" mice. For each dataset we randomly choose Y and C_1, and leave out the observed values of Y for $C_1 = 1$. Figure 3b shows a boxplot of the L_2 loss of the predicted Y values with respect to the real values for the baseline and our method. Given the small size of the datasets, this is a very challenging problem. In this case, our method abstains from making a prediction for 170 cases out of 1000 but performs similarly to the baseline on the remaining cases.

4 Discussion and Conclusion

We have defined a general class of causal domain adaptation problems and proposed a method that can identify sets of features that lead to transferable predictions. Our assumptions are quite general and in particular do not require the causal graph or the intervention targets to be known. The method gives promising results on simulated data. It is straightforward to extend our method to the cyclic case by making use of the results by Forré and Mooij [2018]. More work remains to be done on the implementation side, for example, scaling up to more variables. Currently, our approach can handle about seven variables on a laptop computer, and with recent advances in exact causal discovery algorithms [e.g., Rantanen et al., 2018], a few more variables would be feasible. For scaling up to dozens of variables, we plan to adapt constraint-based causal discovery algorithms like FCI [Spirtes et al., 2000] to deal with the missing-data aspect of the domain adaptation task. We hope that this work will also inspire further research on the interplay between bias, variance and causality from a statistical learning theory perspective.

References

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij. Theoretical aspects of cyclic structural causal models. *arXiv.org preprint*, arXiv:1611.06221v2 [stat.ME], Aug. 2018. URL `https://arxiv.org/abs/1611.06221v2`.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, (AISTATS-07)*, volume 2 of *Proceedings of Machine Learning Research*, pages 107–114, 2007.

P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. *Clingo* = ASP + control: Extended report. Technical report, University of Potsdam, 2014. URL `http://www.cs.uni-potsdam.de/wv/pdfformat/gekakasc14a.pdf`.

M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2839–2848, 2016.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, (UAI-14)*, pages 340–349, 2014.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Do-calculus when the true graph is unknown. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, pages 395–404, 2015.

S. Magliacane, T. Claassen, and J. M. Mooij. Ancestral causal inference. In *In Proceedings of Advances in Neural Information Processing Systems, (NIPS-16)*, pages 4466–4474, 2016.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory (COLT 2009)*, 2009.

F. Markowetz, S. Grossmann, and R. Spang. Probabilistic soft interventions in conditional Gaussian networks. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, (AISTATS-05)*, pages 214–221, 2005.

J. M. Mooij and T. Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 431–439, 2013.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv.org preprint*, https://arxiv.org/abs/1611.10351v3 [cs.LG], Mar. 2018. URL `https://arxiv.org/abs/1611.10351v3`.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 247–254, 2011.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Quiñonero-Candela, M. Suyiyama, A. Schwaighofer, and N. D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2009.

K. Rantanen, A. Hyttinen, and M. Järvisalo. Learning optimal causal graphs with exact search. In *Proceedings of the 9th International Conference on Probabilistic Graphical Models (PGM 2018)*, volume 72 of *Proceedings of Machine Learning Research*, pages 344–355, 2018.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30: 145–157, 2003.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262, 2012.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

A. Storkey. When training and test sets are different: Characterizing learning transfer. In *Dataset Shift in Machine Learning*, chapter 1, pages 3–28. MIT Press, 2009.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *In Proceedings of Advances in Neural Information Processing Systems (NIPS-08)*, pages 1433–1440, 2008.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, (UAI-01)*, 2001.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 819–827, 2013.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3150–3157, 2015.

Supplementary material for "Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions"

Sara Magliacane
IBM Research*
sara.magliacane@gmail.com

Thijs van Ommen
University of Amsterdam
thijsvanommen@gmail.com

Tom Claassen
Radboud University Nijmegen
tomc@cs.ru.nl

Stephan Bongers
University of Amsterdam
srbongers@gmail.com

Philip Versteeg
University of Amsterdam
p.j.j.p.versteeg@uva.nl

Joris M. Mooij
University of Amsterdam
j.m.mooij@uva.nl

1 A stronger version of the assumption in the main paper

In the main paper we make a non-standard assumption, Assumption 2(ii). Here we prove that this is assumption is a weakened version of two more standard assumptions, i.e. assuming the causal Markov and faithfulness assumptions in the source and target domains separately. Note that assuming these two assumptions instead of Assumption 2(ii) implies we cannot have perfect interventions in the target domain, which is otherwise allowed.

Proposition 1. *Assumption 2(ii), i.e. if $\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{S}$ contains Y but not C_1, then[2]*

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 0] \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 1];$$

is implied by the following assumptions:

(a) the pooled source domains distribution $\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)$ is Markov and faithful to $\mathcal{G}^{\backslash C_1}$, and
(b) the pooled target domains distribution $\mathbb{P}(\boldsymbol{V} \mid C_1 = 1)$ is Markov and faithful to $\mathcal{G}^{\backslash C_1}$,

where $\mathcal{G}^{\backslash C_1}$ denotes the induced subgraph of the causal graph $\mathcal{G}$ on the nodes $\mathcal{V} \setminus \{C_1\}$ (i.e., it is obtained by removing C_1 and all edges involving C_1 from the causal graph $\mathcal{G}$).

Proof. Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S} \subseteq \boldsymbol{V} \setminus \{C_1\}$. By assumption, we have that

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = c] \iff \boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S} \, [\mathcal{G}^{\backslash C_1}]$$

holds for both $c = 0, 1$, which directly gives Assumption 2(ii). □

2 Other proofs

Proposition 2. *(Proposition 1 in the main paper) Under Assumption 2,*

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 0] \iff \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \cup \{C_1\} \iff \boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S} \cup \{C_1\} \, [\mathcal{G}]$$

for subsets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{S} \subseteq \boldsymbol{V}$ such that their union contains Y but not C_1.

---

*Most of the work was performed while at the University of Amsterdam.
[2]Here, with $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [C_1 = 0]$ we mean $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{S} \, [\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)]$, i.e., the conditional independence of $\boldsymbol{A}$ from $\boldsymbol{B}$ given $\boldsymbol{S}$ in the mixture of the source domains $\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)$, and similarly for the target domains.

Proof. First of all, $A \not\perp\!\!\!\perp B \mid S \; [C_1 = 0]$ implies (by definition) $A \not\perp\!\!\!\perp B \mid S \cup \{C_1\}$. Second, $A \perp\!\!\!\perp B \mid S \; [C_1 = 0]$ implies (by assumption) $A \perp\!\!\!\perp B \mid S \; [C_1 = 1]$, and taken together, we get $A \perp\!\!\!\perp B \mid S \cup \{C_1\}$. By the Markov and faithfulness assumption (Assumption 2(i)), this holds iff $A \perp\!\!\!\perp B \mid S \cup \{C_1\} \; [\mathcal{G}]$. $\qquad\square$

Example 1. *(Example 2 in the main paper) Assume that Assumptions 1 and 2 hold for two context variables* C_1, C_2 *and three system variables* X_1, X_2, X_3 *with* $Y := X_2$. *If the following conditional (in)dependences all hold in the source domains:*

$$C_2 \perp\!\!\!\perp X_2 \mid X_1 \; [C_1 = 0], \qquad C_2 \not\perp\!\!\!\perp X_2 \mid \emptyset \; [C_1 = 0], \qquad C_2 \perp\!\!\!\perp X_3 \mid X_2 \; [C_1 = 0], \tag{1}$$

then $C_1 \perp\!\!\!\perp X_2 \mid X_1 \; [\mathcal{G}]$, *i.e.,* $\{X_1\}$ *is a separating set for* C_1 *and* X_2.

Proof. In the JCI setting, we assume that in the full ADMG $\mathcal{G}$ over variables $\{C_1, C_2, X_1, X_2, X_3\}$, C_1 and C_2 are confounded and not caused by system variables X_1, X_2, X_3. Furthermore, no pair of system variable and context variables is confounded.

In the context $[C_1 = 0]$, if the conditional independences $C_2 \perp\!\!\!\perp X_2 \mid X_1 \; [C_1 = 0]$ and $C_2 \not\perp\!\!\!\perp X_2 \mid \emptyset \; [C_1 = 0]$ hold, then we can also derive that $C_2 \not\perp\!\!\!\perp X_1 \mid \emptyset \; [C_1 = 0]$, for example using Rule (9) from Magliacane et al. [2016]. Moreover, we know that C_2 is not caused by X_1 and X_2, or in other words $X_1 \not\dashrightarrow C_2$ and $X_2 \not\dashrightarrow C_2$. Thus we conclude that (C_2, X_1, X_2) is an LCD triple [Cooper, 1997] in the context $C_1 = 0$. Since in addition, in this case C_2 and X_1 are unconfounded, the marginal ADMG $\mathcal{G}'$ on $\{C_2, X_1, X_2\}$ (in the context $C_1 = 0$, and hence by Proposition 1 in all contexts) must be given by Figure 1a.

Therefore, the extended marginal ADMG $\mathcal{G}''$ on variables $\{C_1, C_2, X_1, X_2\}$ must also have a directed path from C_2 to X_1 and from X_1 to X_2. C_1 cannot be on these paths, as none of the variables causes C_1, and therefore $\mathcal{G}''$ also contains the directed edges $C_2 \to X_1$ and $X_1 \to X_2$. Moreover, $\mathcal{G}''$ cannot contain any edge between C_2 and X_2, nor a bidirected edge between X_1 and X_2, because that would violate the conditional independence. By construction, in the JCI setting there is a bidirected edge between C_1 and C_2, and that is the only bidirected edge connecting to C_1 or C_2. As we assumed there is no direct effect of C_1 on target X_2, there is no edge between C_1 and X_2 in $\mathcal{G}''$. There is also no directed edge $X_1 \to C_1$ in $\mathcal{G}''$, as the JCI assumption implies none of the other variables causes C_1. Therefore, the marginal ADMG $\mathcal{G}''$ is given by Figure 1b, either with the directed edge $C_1 \to X_1$ present, or without that edge.

If it additionally holds that $C_2 \perp\!\!\!\perp X_3 \mid X_2 \; [C_1 = 0]$, we have two possibilities:

1. if $C_2 \perp\!\!\!\perp X_3 \mid \emptyset \; [C_1 = 0]$ holds, then X_3 is not caused by C_2. This means it cannot be on any directed path from C_2 to X_1, from X_1 to X_2, or be a descendant of X_2. Therefore the full ADMG $\mathcal{G}$ also necessarily contains the directed edges $C_2 \to X_1$ and $X_1 \to X_2$.

2. if $C_2 \not\perp\!\!\!\perp X_3 \mid \emptyset \; [C_1 = 0]$ holds, then in conjunction with $C_2 \perp\!\!\!\perp X_3 \mid X_2 \; [C_1 = 0]$ we can derive $X_2 \dashrightarrow X_3$, for example using Rule (5) from [Magliacane et al., 2016]. This means X_3 must be a descendant of X_2 in the full ADMG $\mathcal{G}$, which implies it cannot be on the directed path from C_2 to X_1, or on the one from X_1 to X_2. Therefore the full ADMG $\mathcal{G}$ also necessarily contains the directed edges $C_2 \to X_1$ and $X_1 \to X_2$.

(a) Marginal ADMG $\mathcal{G}'$.

(b) Set of candidate marginal ADMGs $\mathcal{G}''$.
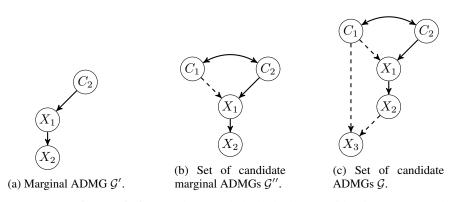
(c) Set of candidate ADMGs $\mathcal{G}$.

Figure 1: ADMGs for proof of Example 1. Each dashed edge can either be present or absent.

Because of the independence statements and JCI assumptions, there cannot be a bidirected edge between X_3 and X_1, X_2, C_1 or C_2. Similarly, there cannot be directed edges from X_3 to one of those nodes. The edges $X_1 \rightarrow X_3$ and $C_2 \rightarrow X_3$ must also be absent.

In both cases, there can be a directed edge from C_1 to X_3. Therefore, the full ADMG $\mathcal{G}$ is of the form given in Figure 1c. In all cases we see that $C_1 \perp X_2 \mid X_1 \, [\mathcal{G}]$, and we conclude that $\{X_1\}$ is a valid separating set.

If the ADMG is as in Figure 2, then a standard feature selection method would asymptotically prefer the subset $\{X_1, X_3\}$ to predict X_2 over the subset $\{X_1\}$ (note that the Markov blanket of X_2 in context $[C_1 = 0]$ is $\{X_1, X_3\}$). As a result, any prediction method trained on all available features using source domain data (i.e., in context $[C_1 = 0]$) may incur a possibly unbounded prediction error when used to predict X_2 in the target domain $[C_1 = 1]$ (for example, if X_3 is an almost deterministic copy of X_2 if $C_1 = 0$, but has a drastically different distribution if $C_1 = 1$). $\qquad\square$
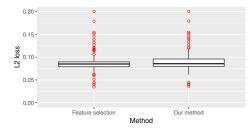
3   Additional results on synthetic data

We provide more information and experimental results for the synthetic data. We adapted the simulator of Hyttinen et al. [2014] to our setting. We generate randomly 200 acyclic models with three system variables, two context variables, and at most two latent variables (chosen randomly, so that the number of latent variables equals 1 or 2 each with probability $1/4$, and 0 otherwise). Each latent variable has two system variables as children, while the other variables have a random number of system variables as children, where system variables must be consistent with a chosen topological ordering, and where we enforce that a context variable may not simultaneously affect all system variables. The system and latent variables are each described by a linear structural equation with independent noise terms distributed as $\mathcal{N}(0, 0.0064)$. In these equations, each variable is multiplied by a coefficient sampled from $\mathcal{N}(0.2, 0.64)$ or $\mathcal{N}(-0.2, 0.64)$ (each with probability $1/2$ per variable). The context variables each correspond to an experimental domain; in their domain, that variable equals 1, otherwise it equals 0. This way, we simulate soft interventions. In order to scale the effect of these interventions, we multiply the coefficients of the context variables by the parameter γ, varying it from 0.1 to 100. We sample N data points each for the observational and two experimental domains. Moreover, we randomly select C_1 and Y from context and system variables respectively. We disallow direct effects of C_1 on Y.
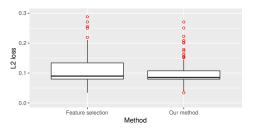
As expected, our method performs well when the target distribution is significantly different from the source distributions. Figure 2 shows different settings with different scales of intervention effects. (In most graphs, the vertical axis has been adjusted to clearly show the boxplot, but leaving out the larger outliers.) In Figure 2a the intervention effects are all scaled by 0.1, resulting in very similar distributions in all domains. In this case, using our method does not offer any advantage with respect to the baseline and it actually performs worse. In the other cases, using our method starts to pay off in terms of prediction accuracy, and the difference increases with the scale of the interventions, as seen in Figure 2d.

In Figure 3, we vary the number of samples N for each regime. The results improve with more samples, especially for our method, since the quality of the conditional independence test improves, but also for the baseline. In particular, as shown in Figure 3a, the accuracy is low for $N = 100$ samples, but it improves substantially with $N = 1000$ samples (Figure 2b).
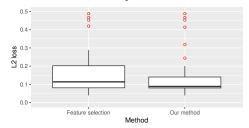
References

G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1(2):203–224, 1997.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, (UAI-14)*, pages 340–349, 2014.

S. Magliacane, T. Claassen, and J. M. Mooij. Ancestral causal inference. In *In Proceedings of Advances in Neural Information Processing Systems, (NIPS-16)*, pages 4466–4474, 2016.
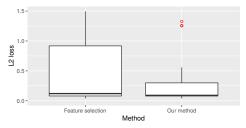
(a) Synthetic data with a small perturbation ($\gamma = 0.1$) and $N = 1000$ samples.

(b) Synthetic data with a medium perturbation ($\gamma = 1$) and $N = 1000$ samples.

(c) Synthetic data with a large perturbation ($\gamma = 10$) and $N = 1000$ samples.
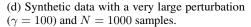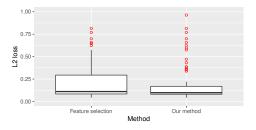
(d) Synthetic data with a very large perturbation ($\gamma = 100$) and $N = 1000$ samples.
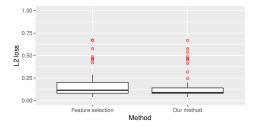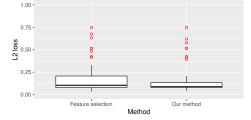
Figure 2: Additional results when varying the causal effect of all interventions (γ).

(a) Synthetic data with $N = 100$ samples per regime and a large perturbation ($\gamma = 10$).

(b) Synthetic data with $N = 1000$ samples per regime and a large perturbation ($\gamma = 10$).

(c) Synthetic data with $N = 5000$ samples per regime and a large perturbation ($\gamma = 10$).

Figure 3: Additional results when varying the sample size per regime (N).