Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

Sara Magliacane[a], Thijs van Ommen[b], Tom Claassen[c], Stephan Bongers[b], Philip Versteeg[b] and Joris M. Mooij[b]

[a] IBM Research, New York [b] Institute for Informatics, University of Amsterdam, The Netherlands [c] Radboud University Nijmegen, The Netherlands

Abstract

In both domain adaptation and causal inference, an important goal is to make accurate predictions in an unseen target domain, where the distribution is different from the source domain(s). We consider **causal domain adaptation** problems, where the domains correspond to different interventions of a single system. The approach we propose exploits causal inference and does not rely on prior knowledge of the causal graph, or of intervention types/targets.

Problem setting: Causal domain adaptation

Unsupervised multi-source domain adaptation with an underlying causal graph, potentially with latent confounders (ADMG).

Domain 1 (observational: wildtype mice)

X_1	X_2	X_3
0.1	0.2	0.5
0.13	0.21	0.49
0.23	0.21	0.51

Measurements of mouse phenotypes:
X_1: red blood cell volume
X_2: platelet count
X_3: white blood cell concentration

Domain 2 (interventional: gene A knocked out)

X_1	X_2	X_3
0.2	0.22	0.92
0.23	0.21	0.99

Domain 3 (interventional: gene B knocked out)

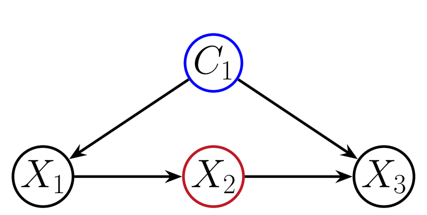
X_1	X_2	X_3
0.5	0.19	?
0.61	0.18	?

X_3 always missing in domain 3: to be predicted

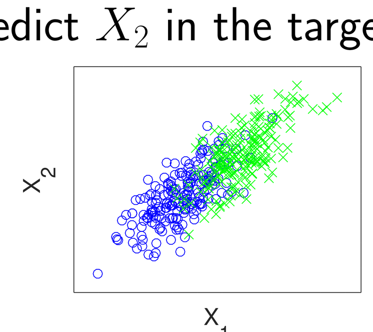
Task: Predict the missing values (all values of X_3 in domain 3).

Example: Standard prediction methods fail

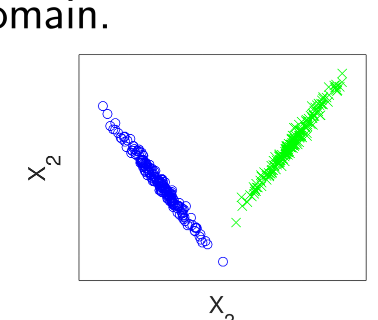
Example: observational source domain ($C_1 = 0$) and interventional target domain ($C_1 = 1$), predict X_2 in the target domain.



True causal ADMG $\mathcal{G}$ ($\mathbb{P}(X_1)$ and $\mathbb{P}(X_3 \mid X_2)$ change from source to target domain, but $\mathbb{P}(X_2 \mid X_1)$ does not)



$\mathbb{P}(X_2 \mid X_1, C_1 = 0) = \mathbb{P}(X_2 \mid X_1, C_1 = 1)$



$\mathbb{P}(X_2 \mid X_3, C_1 = 0) \neq \mathbb{P}(X_2 \mid X_3, C_1 = 1)$

target: *source*:

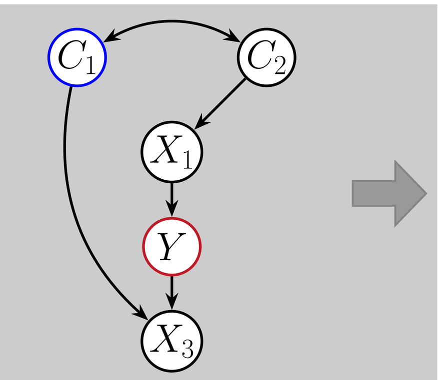
$X_2 \perp_{\mathcal{G}} C_1 \mid X_1$ (*$\{X_1\}$ is separating*) $\implies \mathbb{P}(X_2 \mid X_1, C_1 = 1) = \mathbb{P}(X_2 \mid X_1, C_1 = 0)$

$X_2 \not\perp_{\mathcal{G}} C_1 \mid X_3$ (*$\{X_3\}$ is not*) $\implies \mathbb{P}(X_2 \mid X_3, C_1 = 1) \neq \mathbb{P}(X_2 \mid X_3, C_1 = 0)$

☞ Predictions of X_2 using feature set $\{X_1\}$ (with any regression method) would transfer from source to target domain, because $\{X_1\}$ is a **separating** set of features.

☞ Standard feature selection (applied on the source domain, $C_1 = 0$) would select $\{X_3\}$ or $\{X_1, X_3\}$ as good sets of features for predicting X_2, leading to **arbitrarily large generalization error** (for $C_1 = 1$).

Example of our approach



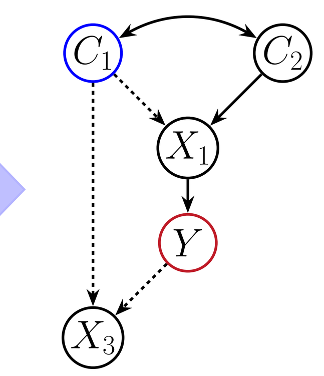
Unknown true ADMG $\mathcal{G}$

C_1	C_2	X_1	Y	X_3
0	0	data	data	data
0	1	data	data	data
1	0	data	?	data

Data with Y missing in target domain

$C_2 \not\perp Y \mid \emptyset \; [C_1 = 0]$
$C_2 \perp Y \mid X_1 \; [C_1 = 0]$
$C_2 \perp X_3 \mid Y \; [C_1 = 0]$
JCI assumptions
$C_1 \not\to_{\mathcal{G}} Y$

Input to causal inference method



Sufficiently reconstructed $\mathcal{G}$

$C_1 \perp Y \mid X_1$

Conclusion: $\{X_1\}$ is separating

C_1	C_2	X_1	Y	X_3
0	0	data	data	data
0	1	data	data	data
1	0	data	!	data

Predict Y using feature set $\{X_1\}$

Overview of our approach

Use conditional independences that can be tested on the available data, to infer enough about the unknown causal graph $\mathcal{G}$ to find **separating** sets $\boldsymbol{A}$ of features ($C_1 \perp_{\mathcal{G}} Y \mid \boldsymbol{A}$). Predictions using such feature sets will transfer across domains, while other predictions may suffer arbitrarily large loss when transferred.

Challenges:

- Types and targets of interventions are also unknown
- Data for Y are consistently missing when $C_1 = 1$, so we cannot test for certain independences, including $C_1 \perp Y \mid \boldsymbol{A}$

Joint Causal Inference (JCI)

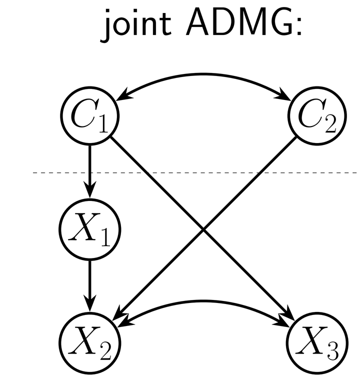
JCI [Mooij, Magliacane and Claassen, 2018] is a meta-algorithm for systematically pooling data from multiple domains, even when intervention types and targets are possibly unknown, reducing causal discovery from different distributions to causal discovery of a single **joint** causal graph with auxilliary **context** variables.

We distinguish:

- **System variables** $\mathcal{X}$, representing the system in each distribution
- **Context variables** $\mathcal{C}$, describing the changes between distributions

context		system		
C_1	C_2	X_1	X_2	X_3
0	0	0.1	0.2	0.5
0	0	0.13	0.21	0.49
0	0	0.23	0.21	0.51
0	1	0.2	0.22	0.92
0	1	0.23	0.21	0.99
1	0	0.5	0.19	?
1	0	0.61	0.18	?

joint ADMG:



JCI assumptions:

1. no system variable directly causes any context variable, and
2. no system variable is confounded with a context variable, and
3. each pair of context variables is purely confounded (i.e. $C_i \leftrightarrow C_{i'} \in \mathcal{G} \wedge C_i \to C_{i'} \notin \mathcal{G}$).

Intuition: We are modelling a generic setting in which the experimenter decides on the performed interventions *before* the measurements are performed (or without having access to the measurements).

Transfer assumptions

The following assumptions enable us to **transfer** information from the source domains to the target domain:

(Y denotes the system variable to be predicted)

1. The mixture of all (training and test) distributions is Markov and faithful w.r.t. an ADMG $\mathcal{G}$;
2. 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,
$$\boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S} \; [C_1 = 0] \implies \boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S} \; [C_1 = 1];$$
3. C_1 has no *direct* effect on Y.

Note that assumption 2 holds if both $\mathbb{P}(\boldsymbol{V} \mid C_1 = 0)$ and $\mathbb{P}(\boldsymbol{V} \mid C_1 = 1)$ are Markov and faithful to the subgraph of $\mathcal{G}$ which excludes C_1.

Dealing with missing data

Due to the missing data, some independences cannot be tested. Some of those can be inferred based on our transfer assumptions.

For an independence $\boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S}$,

- *If $Y \notin \boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{S}$:* independence is testable in data
- *If $Y \in \boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{S}$ and $C_1 \in \boldsymbol{S}$:* follows from the transfer assumptions,
$$\boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{S} \Leftrightarrow \boldsymbol{A} \perp \boldsymbol{B} \mid (\boldsymbol{S} \setminus \{C_1\}) \; [C_1 = 0]$$
- *If $Y \in \boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{S}$ and $C_1 \notin \boldsymbol{S}$:* is untestable and does not follow from the assumptions, e.g. $Y \perp C_1 \mid \boldsymbol{S}$

Implementation with logic-based causal method

Task: "Is $Y \perp C_1 \mid \boldsymbol{A}$?" given all available conditional independences

We modify the method by [Hyttinen, Eberhardt and Järvisalo, 2014] combined with the scores from [Magliacane, Claassen, Mooij, 2016]:

- Input: list of weighted conditional (in)dependence statements (in our case: some are missing)
- **Input statement**: $Y \perp C_1 \mid \boldsymbol{A}$ is true (or false)
- Output: a measure of confidence that the statement is true (or false)

Causal domain adaptation algorithm

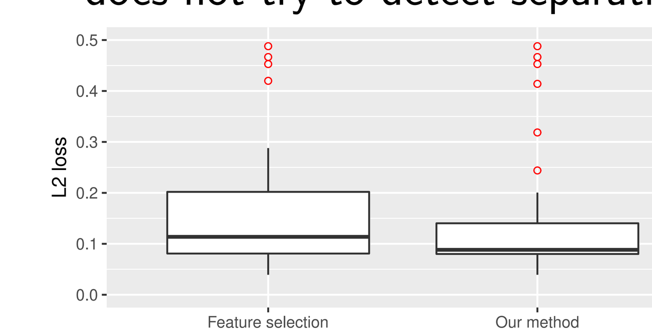
A brute-force strategy to select the feature set $\boldsymbol{A}$ with the best asymptotic guarantee on the prediction error:

for all $\boldsymbol{A} \subseteq (\mathcal{X} \cup \mathcal{C}) \setminus \{Y, C_1\}$ **do**
  $L_{\boldsymbol{A}} \leftarrow$ **estimate of generalization error** when using $\boldsymbol{A}$ to predict Y in source domain
end for
for all $\boldsymbol{A} \subseteq (\mathcal{X} \cup \mathcal{C}) \setminus \{Y, C_1\}$, in increasing order of $L_{\boldsymbol{A}}$ **do**
  if we can **infer that** $Y \perp C_1 \mid \boldsymbol{A}$ **then**
    return predict Y in the target domain using feature set $\boldsymbol{A}$
  end if
end for
return abstain from making a prediction

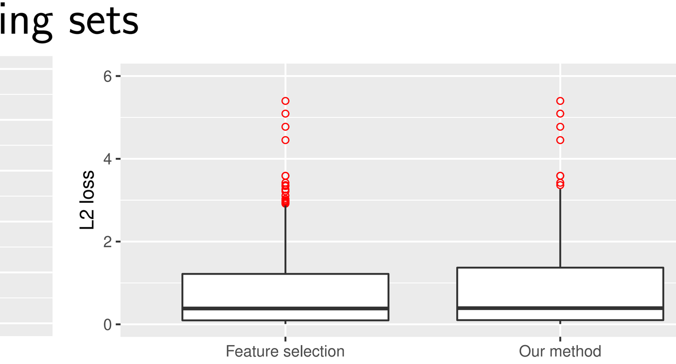
Experimental results

We evaluated our method on simulated and real-world data.

- Simulated data: from randomly generated causal graphs
- Real-world data: hematology data from CRM Causal Inference Challenge (phenotype data for wild-type and single-gene knockout mice)
- In both cases, 2 context and 3 system variables
- Baseline method: Feature selection + regression (random forests); does not try to detect separating sets



Simulated data



Real-world data

Conclusions and future work

- We proposed a method for a class of causal domain adaptation problems, under quite general assumptions and not requiring prior knowledge of causal graph or intervention targets
- Promising results on simulated data, but improvement needed on real-world data
- Scaling up to more variables would also improve prediction quality
- More future work: Study the interplay between bias, variance and causality from a statistical learning theory perspective

References

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.

S. Magliacane, T. Claassen and J. M. Mooij Ancestral causal inference *NIPS 2016*

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI 2014*, pages 340–349.