Joint causal inference on observational and experimental datasets

Sara Magliacane
VU Amsterdam, University of Amsterdam
sara.magliacane@gmail.com

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

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

Abstract

We introduce Joint Causal Inference (JCI), a powerful formulation of causal discovery over multiple datasets in which we jointly learn both the causal structure and targets of interventions from independence test results. While offering many advantages, JCI induces faithfulness violations due to deterministic relations, so we extend a recently proposed constraint-based method to deal with this type of violations. A preliminary evaluation shows the benefits of JCI.

1 Introduction

Answering hypothetical and counterfactual "what if?" questions requires some knowledge about the causal relations of the underlying system. While sometimes these causal relations are known, in many cases we still have to infer them from the available data. Traditionally, causal relations are either recovered from experimental data in which the variable of interest is perturbed, or from observational data, e.g. using the seminal PC/FCI algorithms [16, 19].

Recently, there have been several proposals for combining observational and experimental data to discover causal relations, showing that this combination can improve greatly on the accuracy and identifiability of the predicted causal relations. Some of the proposed methods are *score-based* (e.g., [4, 6]), i.e. they evaluate models using a penalized likelihood score, while others (e.g., [7, 18, 15]) are *constraint-based*, i.e. they use statistical independences to express constraints over possible models. One of the primary advantages of constraint-based methods over score-based methods is their ability to naturally handle latent variables, in particular, confounders.

In [11] we propose Joint Causal Inference (JCI), a formulation of causal discovery over multiple datasets in which we jointly learn both the causal structure and targets of interventions from independence test results. A similar approach was already proposed for score-based methods in [4], but here we extend it to constraint-based methods. Our goal is to unify the idea of joint inference from observational and experimental data from [4] with the advantages that constraint-based methods have over score-based methods, namely, the ability to handle latent confounders naturally, and, especially in the case of logic-based methods, an easy integration of background knowledge.

Existing constraint-based methods for multiple datasets typically learn the causal structure on each dataset separately and then merge the inferred structures. Typically, they support only (perfect) interventions on known targets [7, 18]. Instead, JCI: (1) allows for several different types of interventions and learns intervention targets; (2) systematically pools data across different datasets, which improves the statistical power of independence tests; and (3) improves the identifiability and accuracy of the predicted causal relations.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

R	I_1	I_2	X_1	X_2	X_4
1	20	0	0.1	0.2	0.5
1	20	0	0.13	0.21	0.49
1	20	0	...	...	...
2	20	1	...	...	...
3	30	0	...	...	...
4	30	1	...	...	...
5	30	2	...	...	...

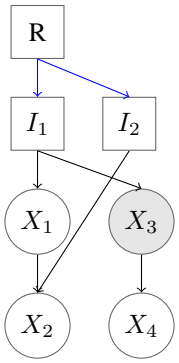


Figure 1: A set of five experimental datasets in a raw data form (left) and as a causal influence diagram representing the causal structure of the system variables $X_1, \ldots, X_4$, the regime variable R and the intervention variables I_1, I_2 (right). The intervention variable I_1 represents the temperature at which each experiment was performed, while the intervention variable I_2 represents the dosage of a drug added in some of the experiments. In the causal diagram, we represent system variables $\{X_j\}_{j\in\mathcal{X}}$ as circles, which are filled for hidden variables, while we represent the regime R and the intervention variables I_1, I_2 as squares.

JCI is a challenge for current constraint-based methods because of their susceptibility to violations of the *Causal Faithfulness assumption*. Specifically, it induces faithfulness violations due to deterministic relations between intervention variables, which cannot be handled by standard constraint-based methods. A simple example of this type of violation is shown in the graph in Figure 1 where $X_1 \perp\!\!\!\perp I_1 \mid R$, because I_1 is determined by R. We propose a simple but effective strategy for dealing with this type of faithfulness violations. We implement it in ACID, a determinism-tolerant extension of ACI [10], a recently proposed logic-based causal discovery method that improves reliability of the output by exploiting redundant information in the data. In our preliminary evaluation on synthetic data we show that JCI with ACID improves on the accuracy of the causal predictions with respect to simply merging separately learned causal graphs.

2 Joint Causal Inference (JCI)

We propose to model jointly in a single causal graph several observational or experimental datasets $\{D_r\}_{r\in\{1\ldots n\}}$, each representing the data after a (possibly empty) set of interventions on one or more, possibly unknown, targets. We assume that there is a unique underlying causal DAG $\mathcal{G}$ in all of these datasets, defined over the same set of variables that we call the *system variables*, $\{X_j\}_{j\in\mathcal{X}}$, some of which are possibly hidden. This assumption precludes certain types of interventions, notably, perfect interventions [14]. On the other hand, it allows for many other types of interventions, e.g., soft interventions [12], mechanism changes [17], fat-hand interventions [4], activity interventions [13], etc., as long as they do not induce new (in)dependences, which can be seen as modifications to the underlying DAG.

Each dataset D_r has an associated joint probability distribution $P_r((X_j)_{j\in\mathcal{X}})$. Using the terminology from [3], we call the different distributions in the datasets *regimes*. In related work different names have been used, e.g. *experimental conditions* or *environments* [15]. We introduce two types of dummy variables in the data:

- a **regime variable** R, representing which dataset D_r a data point is from, i.e., $\forall r = 1 \ldots n$, $R = r$ for data from D_r.
- **intervention variables** $\{I_i\}_{i\in\mathcal{I}}$, which are functions of the regime R. Intuitively, intervention variables represent the interventions performed in each dataset. We describe an example in Figure 1. In absence of any information on the interventions performed in the datasets, we can use as intervention variables the indicator variables for each of the datasets.

We assume $\mathcal{G}$ with the introduced dummy variables can be represented as an acyclic Structural Causal Model (SCM) with independent exogenous variables $\{E_k\}_{k\in\mathcal{X}\cup\{R\}}$:

$$\begin{aligned} P((E_k)_{k\in\mathcal{X}\cup\{R\}}) &= \prod_{k\in\mathcal{X}\cup\{R\}} P(E_k) \\ R &= E_R \\ \forall i \in \mathcal{I} : I_i &= g_i(R) \\ \forall j \in \mathcal{X} : X_j &= f_j(X_{pa(X_j)}, I_{ipa(X_j)}, E_j) \end{aligned} \tag{1}$$

Here, $pa(X_j)$ are the system variable parents of j, $ipa(X_j)$ the intervention parents and E_j is the exogenous parent of X_j.

We represent this SCM with a causal influence diagram $\mathcal{C}$, see e.g. [3], and assume the Causal Markov and Minimality assumptions hold in $\mathcal{C}$. We show an example of a causal influence diagram in Figure 1, where we model five datasets with the same underlying causal graph.

Intervention variables are functions of the regime variable, and do not have any associated noise. This means that they are *determined* by the regime. In general, other deterministic relations may arise. For example, consider an intervention variable that is an indicator of whether the regime is an odd number, and another which indicates if the regime is an even number. These two variables determine each other, even if it is not clear from the causal influence diagram.

In this extended abstract, we do not consider the general case, and assume that the only deterministic relations are the regime R determining each of the intervention variables $\{I_i\}_{i\in\mathcal{I}}$. For this restricted case of *functionally determined* relations, defined recursively as variables that are fully determined by their parents, Geiger et al. [5] proved that D-separation is sound and complete under the Causal Markov and Minimality assumption. For the more general case, an extension of D-separation is presented in [16], which retains completeness for the restricted case, but is not proven to be complete in general. We conjecture that for the more general case of deterministic relations between dummy variables, D-separation is complete, but we leave the completeness proof for future work.

We are interested in the completeness of D-separation, because we wish to use it to relax the standard Causal Faithfulness assumption. In our setting this assumption is too restrictive, so we relax it to allow for violations due to deterministic relations between the regime and the intervention variables. We define our relaxed version, that we call *D-Faithfulness assumption*, as follows: for three disjoint sets of variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}$ and a probability distribution P that satisfies both the Causal Markov assumption for $\mathcal{C}$ and the list of deterministic relations $\mathcal{D}$, we assume that $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{W}\ [P] \implies \boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{W}\ [\mathcal{D}, \mathcal{C}]$, where $\perp_D$ represents D-separation as defined in [16].

Finally, we define Joint Causal Inference (JCI) as the problem of reconstructing the causal inference diagram that represents jointly all datasets and intervention variables from independence test results. It can be shown that the ideas behind some previous approaches, e.g., [2, 8, 15], can be seen as special cases of JCI.

3 Extending constraint-based methods for JCI

JCI provides some challenges for current constraint-based methods:

- faithfulness violations due to deterministic relations between the dummy variables,
- the availability of complex background knowledge on the dummy variables that can improve structure learning and recover from some of the faithfulness violations, e.g. R can only cause a system variable through an intervention variable.

There is some work on dealing with faithfulness violations in the PC algorithm [9], but it assumes causal sufficiency (in our context, no hidden variables in $\mathcal{G}$), and cannot handle complex background knowledge. Other constraint-based algorithms, specifically, logic-based causal discovery methods, e.g., [7, 18, 10] can handle complex background knowledge and causal insufficiency, but cannot deal with faithfulness violations due to deterministic relations.

We propose a simple but effective strategy for dealing with faithfulness violations due to deterministic relations. We rephrase the constraints of a constraint-based algorithm in terms of d-separations and d-connections, instead of independence test results. At testing time we decide for each indepen-

dence test result which d-separations and d-connections can be soundly derived from it and provide them as input to the modified constraint-based algorithm. We use the following rules:

- $X \not\perp\!\!\!\perp Y|\boldsymbol{W} \implies X \not\perp_d Y|\boldsymbol{W}$,
- $X \notin \text{DET}(\boldsymbol{W}) \wedge Y \notin \text{DET}(\boldsymbol{W}) \wedge X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \implies X \perp_d Y \,|\, \text{DET}(\boldsymbol{W})$,

where $\perp_d$ is standard *d-separation*, $\not\perp_d$ is d-connection, and $\text{DET}(\boldsymbol{W})$ are the variables determined by (a subset of) $\boldsymbol{W}$. Note that the above procedure outputs d-separations only for a subset of independence test results, ignoring independences when X or $Y \in \text{DET}(\boldsymbol{W})$. It can be shown that we can get sound d-separations and d-connections using this procedure, under the Causal Markov, Minimality and D-Faithfulness assumptions.

This simple strategy can be applied to any constraint-based method, providing that it can deal with partial inputs, i.e. missing results for a certain independence test. Logic-based methods as [7, 10] can be run out-of-the-box with partial inputs, while other standard algorithms like FCI [19] would require possibly complicated extensions. Anytime FCI [1] allows one to ignore (in)dependences above a certain order, but up to that order they should all be available.

4 Ancestral Causal Inference With Determinism (ACID)

We implement the proposed strategy in ACID (Ancestral Causal Inference with Determinism) as a determinism-tolerant extension of ACI (Ancestral Causal Inference) [10]. ACI is a recently-introduced logic-based causal discovery method that accurately reconstructs ancestral structures ("indirect" causal relations), also in the presence of latent variables and statistical errors. Moreover, it provides a method for scoring the reliability of causal predictions, which roughly approximates their marginal probability. For brevity, in this extended abstract we only provide a few examples of the extension from ACI to ACID.

ACI is based on a set of logical rules. For example, for variables X, Y and a set of variables $\boldsymbol{Z}$, where $X \not\dashrightarrow \boldsymbol{Z}$ represents the fact that X does not cause any of the variables in set $\boldsymbol{Z}$:

$$(X \perp\!\!\!\perp Y \mid \boldsymbol{Z}) \wedge (X \not\dashrightarrow \boldsymbol{Z}) \implies X \not\dashrightarrow Y. \tag{2}$$

ACID reformulates the logical rules of ACI in terms of *d-separation*. This reformulation completely decouples the rules from any assumption on the relation between (in)dependences and d-separations/d-connections, e.g., Causal Faithfulness. For example, the above rule can be simply rewritten as:

$$(X \perp_d Y \mid \boldsymbol{Z}) \wedge (X \not\dashrightarrow \boldsymbol{Z}) \implies X \not\dashrightarrow Y. \tag{3}$$

The new rules can be then used with the procedure for deriving d-separations from independence test results described in the previous Section. To improve the identifiability and accuracy of the predictions, we also add as background knowledge a series of logical rules. For brevity, we omit the complete list of rules, and just show a simple example:

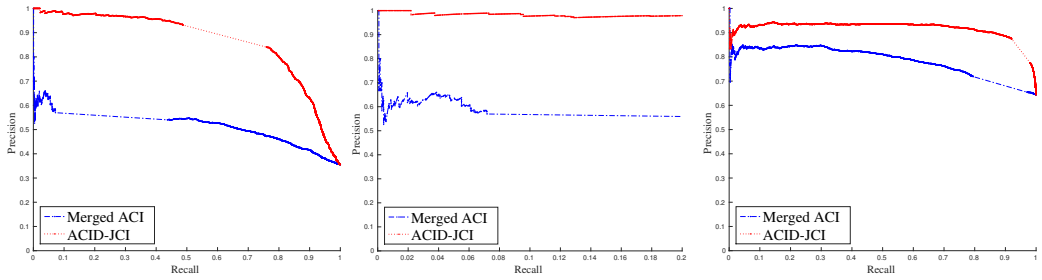
$$\forall i \in \mathcal{I}, \forall j \in \mathcal{X} : (X_j \not\dashrightarrow R) \wedge (X_j \not\dashrightarrow I_i).$$

This rule expresses the fact that the regime variable is, by definition, never caused by any other variable. Adding this and other background knowledge rules provides a simple means to ruling out several spurious candidate causal structures, showcasing a main advantage of logic-based causal discovery methods.

5 Preliminary evaluation

We run a preliminary evaluation of ACID on 600 randomly generated datasets. Each dataset contains one observational and three experimental regimes for a causal system with four system variables. The simulator builds up on the simulator for linear acyclic models with latent variables and Gaussian noise described in [7, 10] and implements soft interventions on unknown targets. This constrains the comparison with many constraint-based methods like [7, 18].

In our preliminary evalution we compare the ancestral structure ("indirect" causal relations) predicted by ACID with a naive baseline, in which we merge ancestral structures learned on each



(a) PR ancestral (b) PR ancestral (zoom) (c) PR nonancestral

Figure 2: Results on synthetic data sets showing the potential gain in precision and quality that can be obtained via JCI. The left column shows the precision-recall (PR) curve for ancestral predictions, the middle column shows a zoomed-in version in the interval (0,0.02), while the right column shows the PR curve for nonancestral predictions.

dataset separately with ACI. As inputs to both algorithms we provide the same weighted independence test results, computed with a test based on partial correlation and Fisher's z-transform with significance threshold $\alpha = 0.05$, and weighted using the frequentist weighting scheme from [10]. In Figure 2 we report the precision and recall (PR) curves for predicting ancestral relations ("indirect" causal relations) and nonancestral relations (the absence of such a causal relation). We can see from the figure that ACID improves significantly on the accuracy of the predictions with respect to the baseline. Although limited, the preliminary results are quite promising.

6 Conclusions and future work

In this extended abstract, we briefly presented Joint Causal Inference (JCI), a powerful formulation of causal discovery over multiple datasets that was previously unexploited by constraint-based methods. Current constraint-based methods cannot be applied to JCI because of faitfulness violations, so we proposed a simple strategy for dealing with this type of faithfulness violations, and showed some preliminary results on its performance. For the full story, we refer the reader to [11].

In future work, we plan to investigate other possible strategies or extensions to existing algorithms for dealing with faithfulness violations. Moreover, we plan to improve on the preliminary evaluation by comparing with more state-of-the-art algorithms and also on different tasks, e.g. learning intervention targets, both on synthetic and real-world data. Finally, although very accurate and flexible, logic-based methods as [7, 10] are limited in the number of possible variables they can handle. JCI introduces additional variables, reducing their scalability. We plan to investigate improvements to the execution times of methods like ACID.

Acknowledgments

SM, JMM and TC were supported by NWO, the Netherlands Organization for Scientific Research (VIDI grant 639.072.410). SM was also supported by the Dutch programme COMMIT/ under the Data2Semantics project. TC was also supported by NWO grant 612.001.202 (MoCoCaDi), and EU-FP7 grant agreement n.603016 (MATRICS).

References

[1] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[2] G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Min. Knowl. Discov.*, 1(2):203–224, Jan. 1997.

[3] A. P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

[4] D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 107–114, 2007.

[5] D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20(5):507–534, 1990.

[6] A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

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

[8] V. Lagani, I. Tsamardinos, and S. Triantafillou. Learning from mixture of experimental data: A constraint–based approach. In *Artificial Intelligence: Theories and Applications: 7th Hellenic Conference on AI, SETN 2012, Lamia, Greece, May 28-31, 2012. Proceedings*, 2012.

[9] J. Lemeire, S. Meganck, F. Cartella, and T. Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *Int. J. Approx. Reasoning*, 53(9):1305–1325, Dec. 2012.

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

[11] S. Magliacane, T. Claassen, and J. M. Mooij. Joint causal inference on observational and interventional datasets. *arXiv.org preprint*, arXiv:1611.10351 [cs.LG], Nov. 2016.

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

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

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

[15] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 2015.

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

[17] J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–521, 2001.

[18] S. Triantafillou and I. Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

[19] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17):1873–1896, Nov. 2008.