Data-driven covariate selection for nonparametric estimation of causal effects

Doris Entner*, **Patrik O. Hoyer***, **Peter Spirtes****
*HIIT and Department of Computer Science, University of Helsinki
**Department of Philosophy, Carnegie Mellon University

Abstract

The estimation of causal effects from non-experimental data is a fundamental problem in many fields of science. One of the main obstacles concerns confounding by observed or latent covariates, an issue which is typically tackled by adjusting for some set of observed covariates. In this contribution, we analyze the problem of inferring whether a given variable has a causal effect on another and, if it does, inferring an adjustment set of covariates that yields a consistent and unbiased estimator of this effect, based on the (conditional) independence and dependence relationships among the observed variables. We provide two elementary rules that we show to be both sound and complete for this task, and compare the performance of a straightforward application of these rules with standard alternative procedures for selecting adjustment sets.

1 INTRODUCTION

In many fields of science researchers are interested in estimating the causal effect of one variable on another. For instance, in epidemiology one might study the effect of a given treatment (or exposure) on the health of the patient, while in economics one may be interested in the influence of education on the income level of individuals. When possible, the preferred means of estimating such effects is using randomized controlled experiments, in which the 'treatment' variable (the cause) is randomized while the 'outcome' variable (the effect) is passively observed. Unfortunately, this technique is not possible in many cases. In the examples given above, it would be unethical to deliberately expose individuals to harmful substances, and it would be in practice impossible to actively control the level of education that individuals achieve.

When randomized experiments are not available, researchers are left with no option but to attempt to infer from non-experimental ('passive observational') data whether a variable has an effect on another, and, if it does, the magnitude of the causal effect. In such studies, one of the main risks is bias due to confounding by observed or unobserved covariates: A variable that has an effect on both the treatment and the outcome variables may significantly bias the estimated effect unless this confounding is properly accounted for. Thus, typically, some covariates are measured and 'adjusted for' (also termed 'controlled for') when estimating the causal effect of the treatment on the outcome.

If the full causal structure among all the variables is known, there are algorithms to determine whether there exists a consistent and unbiased estimator of the desired causal effect (Pearl, 2009; Shpitser et al., 2010; Shpitser and Pearl, 2006). However, if the causal structure among the variables is *not* known, it is not clear how to select the set of covariates to adjust for. Some investigators suggest adjusting for *all* measured covariates, while others propose adjusting only for those with certain statistical properties, such as the ones associated with both the treatment and the outcome. A recent study (VanderWeele and Shpitser, 2011) advocated adjusting for those covariates that are known to influence the treatment or the outcome, or both. Many of the proposed approaches are mutually contradictory in their prescriptions. Thus, when the causal structure among the variables is not known the problem is still largely unsolved, both in terms of the theoretical methodology and in terms of practical deployment of known results. In this paper, we provide a set of simple rules that, under well specified conditions, identify the appropriate set of covariates to adjust for, based on testable (conditional) dependence and independence relations among the observed variables.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

2 MODEL AND PROBLEM SETUP

We derive all of our results in the framework of modeling causal relationships using directed graphical models (Spirtes et al., 2000; Pearl, 2009). The *causal structure* of the underlying system is represented by a directed acyclic graph (DAG) over a set $\mathcal{V} = \{x, y\} \cup \mathcal{W} \cup \mathcal{U}$ of random variables, where x represents the 'treatment' variable, y represents the 'outcome' variable, $\mathcal{W}$ consists of the set of observed covariates, and $\mathcal{U}$ represents an unknown set of latent (unobserved) covariates, assuring that the full set of variables $\mathcal{V}$ forms a causally sufficient set, i.e. any common cause of two or more variables in $\mathcal{V}$ is in $\mathcal{V}$. Note that $\mathcal{W}$ and $\mathcal{U}$ are allowed to be empty.

We assume that y is not a causal ancestor of x, and that x and y are not causal ancestors of any variables in $\mathcal{W}$. This represents the situation where the covariates correspond to 'background conditions' that may influence both the treatment and the effect but may not be influenced by these, and it is further known that the effect cannot influence the treatment. Such an assumption is typically reasonable in cases where a clear time ordering exists among $\mathcal{W}$, x, and y. Examples are given in Figure 1.

We further make the common assumption of *faithfulness*, i.e. all independencies in the distribution $P(\mathcal{V})$ are due to the structure of the graph, rather than the specific form of the relationships among the variables. Under this assumption, d-separation (Pearl, 2009, stated below) in the graph corresponds precisely to conditional independence in the distribution $P(\mathcal{V})$.

Definition 1 (D-separation)**.** *A path p is* blocked *by a set $\mathcal{Z}$ if (a) p contains a chain $v_i \to v_k \to v_j$ or a fork $v_i \leftarrow v_k \to v_j$ with $v_k \in \mathcal{Z}$, or (b) p contains a* collider *$v_i \to v_k \leftarrow v_j$ such that neither v_k nor any of its descendants are in $\mathcal{Z}$. If a path is not blocked it is called* active*. Variables v_1, v_2 are* d-separated *by a set $\mathcal{Z}$ if every path between v_1 and v_2 is blocked by $\mathcal{Z}$.*

Note that we make no other assumptions on the functional form of the relationships, or the distributions involved. We do assume that there is no selection bias, i.e. whether or not a unit is included in the sample is not affected by any variable causally related to any $v \in \mathcal{V}$. Typically this is achieved by a random sample.

The problem is to infer if x has a causal effect on y, and if it does, to obtain a consistent and unbiased estimator for the causal effect of x on y, i.e. $P(y \,|\, \mathrm{do}(x))$ in the notation of Pearl (2009). This may or may not be possible by "adjusting" for an appropriate set $\mathcal{Z} \subseteq \mathcal{W}$. When the DAG over all variables $\mathcal{V}$ (observed and latent) is known, the following theorem due to Pearl (2009) specifies which sets $\mathcal{Z}$ are appropriate.

Theorem 1. ***('Back-Door Adjustment')*** *If a set of variables $\mathcal{Z} \subseteq \mathcal{W}$ satisfies the* back-door criterion *relative to (x, y), i.e.*

(i) no node in $\mathcal{Z}$ is a descendant of x; and

(ii) $\mathcal{Z}$ blocks every back-door path from x to y (i.e. paths from x to y containing an arrow into x: $x \leftarrow \ldots \to y$),

then the causal effect of x on y is identifiable and is given by

$$P(y \,|\, \mathrm{do}(x)) = \sum_{Z} P(y \,|\, x, Z) P(Z). \qquad (1)$$

In our setting, condition (i) is always fulfilled. Thus, if $\mathcal{Z}$ d-separates x from y in the model in which the edge from x to y is cut (if it exists), then $\mathcal{Z}$ satisfies the criterion and is termed *admissible*. It is further known that for any set $\mathcal{Z}$ which is *not* admissible, there exist models in which adjusting for $\mathcal{Z}$ yields an inconsistent and biased estimator of the desired causal effect. In this sense, admissibility is exactly the property that is sought (Shpitser et al., 2010).

With this criterion, it is easy to see why simple strategies such as adjusting for all covariates, for none of the covariates, or all covariates associated to both x and y, can fail. Consider the two graphs in Figure 1 **(a)** and **(b)**. In **(a)**, $\mathcal{Z} = \{w_1\}$ is an admissible set, whereas in **(b)**, $\mathcal{Z} = \{w_1\}$ is not admissible. On the other hand, $\mathcal{Z} = \emptyset$, is not admissible in **(a)**, but admissible in **(b)**. Thus, all three of these simple strategies can fail. In fact, since these two models entail the same set of independencies among the observed variables (there are none), it is impossible to decide from testable dependencies and independencies alone whether w_1 should be in the adjustment set or not.

Typically, the full causal structure is not known. Nevertheless, in many cases the independencies and dependencies among the observed variables do provide sufficient information to identify the causal effect. Our goal is thus a procedure that, under the stated assumptions, uses (testable) independencies and dependencies in the data to output exactly one of the following:

`'±'`: x has a causal effect on y, and the effect is found by back-door adjustment with a given admissible set $\mathcal{Z} \subseteq \mathcal{W}$

`'0'`: x has no causal effect on y

`'?'`: we do not know whether x has a causal effect on y or not.

At first sight, it might seem that there could be cases in which the data allows us to infer that there is a causal effect of x on y, but without us knowing an admissible adjustment set. We will show that, under the given assumptions, such cases do not occur.

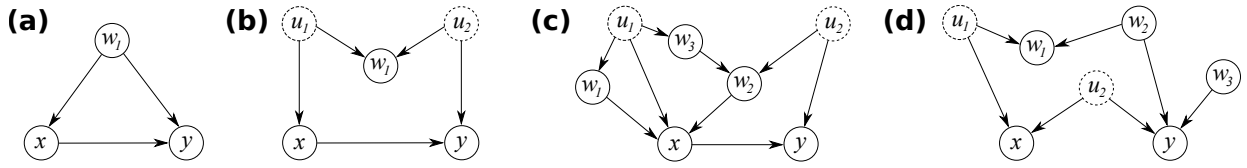


Figure 1: Example graphs with latent variables $u_i \in \mathcal{U}$, and observed covariates $w_j \in \mathcal{W}$, for all i and j.

3 INFERENCE RULES

We introduce the following simple rules for solving the above problem.

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and
(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

then infer `'±'` and give $\mathcal{Z}$ as an admissible set.

R2: If there exists a set $\mathcal{Z} \subseteq \mathcal{W}$ such that

(i) $x \perp\!\!\!\perp y \mid \mathcal{Z}$,

or, if there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(ii) $w \not\perp\!\!\!\perp x \mid \mathcal{Z}$, and
(iii) $w \perp\!\!\!\perp y \mid \mathcal{Z}$,

then infer `'0'`.

If neither R1 nor R2 apply, then we simply output `'?'`.

To get an intuition for why rule R1 is appropriate, consider the following. Condition (i) ensures that there exist (one or more) active paths from w to y, given $\mathcal{Z}$, which, by condition (ii) must all pass through x, since including x in the conditioning set blocks all the paths. This implies that there are (one or more) active paths from w to x. If there existed an active *back-door* path from x to y then condition (ii) could not hold, because then the concatenation of the former with the latter would be an active path, with x as a collider in the conditioning set. One example graph in which R1 applies is given in Figure 1 **(c)**, with $w = w_1$, $\mathcal{Z} = \{w_2, w_3\}$.

The main idea of rule R2 is twofold: First, if there existed a non-zero effect of x on y then, by faithfulness, x and y would be dependent conditional on any subset of the remaining variables, so condition (i) is sufficient on its own to infer a zero effect. Second, conditions (ii) and (iii) of R2 together may allow us to detect a zero effect even in the presence of a latent confounder between x and y: Condition (ii) ensures that there exist (one or more) active paths from w to x given $\mathcal{Z}$, so if there existed an edge from x to y then adding that edge to the previous path would immediately yield an active path from w to y, contradicting condition (iii).

Figure 1 **(d)** gives an example where R2 (ii) and (iii) apply, with $w = w_1$ and $\mathcal{Z} = \{w_2\}$.

Next, we provide the formal results stating that, in the large sample limit, the combination of these two simple rules is both *sound* (i.e. whenever we infer either `'±'` or `'0'`, this is correct) and *complete* (i.e. we only output `'?'` when, based on conditional independencies and dependencies among the observed variables, it cannot be known whether x has a causal effect on y or not).

Theorem 2. *Given an independence oracle, under the assumptions stated in Section 2, whenever rule R1 or R2 applies, the corresponding inference is correct.*

The formal proof of the theorem is given in the Appendix. The main arguments were already discussed above. Note that as long as consistent tests of independence and consistent estimators for parameters exist, the estimated causal effect will be pointwise consistent, following similar arguments as in Spirtes et al. (2000, Ch. 12.4), and Robins et al. (2003).

Theorem 3. *Given an independence oracle, under the assumptions stated in Section 2, whenever neither rule R1 nor R2 applies it is impossible to determine, based on the conditional independencies and dependencies among the observed variables alone, whether x has a causal effect on y or not.*

The proof is given in the Appendix. The main idea is that if neither rule applies using an independence oracle, then there exist causal structures with and without an edge from x to y, which entail the same dependencies and independencies among the observed variables, and hence it is impossible to reliably infer `'±'` or `'0'` solely based on testable independencies.

4 PRACTICAL ALGORITHM

In this section we discuss how the inference rules introduced in the previous section can be used in practice to infer `'±'`, `'0'`, or `'?'`, from finite-sample data.

For small sets of observed covariates it is possible to go through all possible sets $\mathcal{Z}$ and pairs $(w, \mathcal{Z})$, obtaining p-values for the independence tests required by rules R1 and R2. In each test, we must infer either dependence or independence. A conservative approach

is to infer a dependence only when the p-value is below some quite low threshold, while requiring that the p-value is above a different, much higher threshold to infer independence. Note that, strictly speaking, a statistical dependence can never be rejected in such statistical tests, so in practice we must infer independence when the null hypothesis of independence *cannot* be rejected.[1] At present, we also disregard the problem of multiple testing, and that underlying samples are partly shared, so that the statistical tests are correlated. We further do not take into account the power of the test. These are possible extensions of our work.

With the above approach, one may find multiple sets $\mathcal{Z}$ and pairs $(w, \mathcal{Z})$ which satisfy the conditions of the rules. With finite sample data, one often finds cases where *both* rules R1 and R2 apply, contradicting each other. A naïve procedure would be based on majority voting: If the necessary conditions for R1 were found more often than the necessary conditions for R2 (normalized by the number of tests run for each rule), infer '±'; if the reverse holds, infer '0'; and if neither holds, infer '?'. This would, however, not be a very conservative approach. Instead, we use a simple Bayesian classifier based on a separate training set of simulated data, where the correct answer ('±', '0', or '?') was given by d-separation computed on the (known) true causal structure. The input is simply the difference of the frequencies of R1 and R2, and the class-conditional distributions are Gaussian. The main effect is that if the total number of applications of rules R1 and R2 is very low, or the rules apply in comparable frequencies, we infer the uninformative but never incorrect '?'.[2] To ensure that we err on the side of caution, we compare the causal effects estimated when using all admissible sets found with R1. If these are significantly different from each other, we cannot give an estimate of the causal effect and hence output '?'.

When the number of observed covariates is high (several tens of variables or more), it is not possible to consider all sets $\mathcal{Z}$ or pairs $(w, \mathcal{Z})$ as the number of sets $\mathcal{Z}$ grows exponentially in the number of covariates. In this case, one can (a) limit the size of the sets $\mathcal{Z}$ considered, or (b) randomly sample sets $\mathcal{Z}$ and pairs $(w, \mathcal{Z})$ from the full set. With the latter approach, one obtains estimates of the frequencies with which the conditions of rules R1 and R2 apply, and these estimates can be directly used as described above.

[1] This is a fundamental issue with all constraint-based approaches to causal inference relying on faithfulness.

[2] In the infinite sample limit, '?' is only inferred when the structure of the underlying graph makes it impossible to tell whether there exists a causal effect or not. However, in the finite sample case it is prudent to output '?' whenever there isn't sufficient evidence for either '±' or '0'.

5 RELATED WORK

To the best of our knowledge, the existing work most closely related to ours is that of Spirtes and Cooper (1999), and Chen et al. (2007). In both of these papers the authors search for triples (w, x, y) (in our notation), in which w is *exogenous* (there are no edges into w in the generating model, known either from background knowledge or due to randomization of w), w, x, and y are all pairwise marginally dependent, and w is independent of y conditional on x. If such a triple is found, they infer that x has a non-zero causal effect on y, and this effect is unconfounded. Our inference rules generalize this approach by allowing non-empty conditioning sets $\mathcal{Z}$, as well as by replacing the assumption of w being exogenous with the weaker assumption of x not being an ancestor of any w. On the other hand, Spirtes and Cooper (1999), and Chen et al. (2007) do not assume that y is not an ancestor of x, whereas our inference rules currently rely on this assumption.

Two recent papers (de Luna et al., 2011; VanderWeele and Shpitser, 2011) discuss the problem of covariate selection under the same acyclicity and partial ordering assumption as our approach. De Luna et al. (2011) further assume faithfulness, and that a subset of the covariates $\mathcal{W}' \subseteq \mathcal{W}$ is admissible, and this subset $\mathcal{W}'$ is *known* a priori. They suggest a selection procedure, using a series of independence tests, to find a minimal subset of $\mathcal{W}'$ which is still admissible. VanderWeele and Shpitser (2011), on the other hand, only assume that there *exists* some admissible set among all covariates $\mathcal{W}$, which does not have to be a priori known. Their procedure however requires prior knowledge about which covariate $w \in \mathcal{W}$ is a cause (i.e. ancestor in the true graph) of x or of y: They show that the adjustment set $\mathcal{Z}$ including all those covariates w which are a cause of x, or a cause of y, or of both, satisfies the back-door criterion and so is admissible. As $\mathcal{Z}$ might include redundant variables, procedures are also provided for selecting a subset of $\mathcal{Z}$ which is still admissible. While the extra information required by VanderWeele and Shpitser (2011) is somewhat restrictive in terms of applicability of the method, the main drawback of both of the above methods is the assumption that there exists an admissible set among the observed covariates $\mathcal{W}$. If this does not hold, neither of the two methods is able to detect this violation of the assumption, and this can lead to an inconsistent and biased estimator of the causal effect.

For *linear non-Gaussian* acyclic models (each variable is a linear combination of its parents and a non-Gaussian disturbance), Entner et al. (2012) tackle the problem of identifying whether an estimator of the causal effect of x on y is consistent and unbiased, given the same partial ordering assumption as in this paper.

In that approach, non-Gaussianity is essential, as the method fails when applied to linear Gaussian models (such as the ones used in the simulations of this paper).

A more general approach to causal discovery with hidden variables is given by the FCI algorithm (Spirtes et al., 2000), which is a constraint based search method that infers, under the faithfulness assumption, *ancestral* relationships among all observed variables using conditional independence tests. It is straightforward to incorporate background knowledge about a partial ordering among the variables in FCI (Spirtes et al., 2000). However, while Zhang (2008) showed that FCI is sound and complete, it is not known whether completeness still holds when incorporating the above form of background knowledge in the algorithm. Using the output of the FCI algorithm, it is possible to make the inferences '±', '0', or '?'. (The details of the procedure are left to the Supplementary Material as this requires an understanding of ancestral graphs.) Thus, our approach can be seen as a special case of the much more general apparatus of FCI, tailored to the specific circumstances discussed. This not only allows a method that can be understood with only the most basic knowledge of d-separation in DAGs, but also yields a relatively simple proof of completeness. It also does not attempt to infer irrelevant causal features among the covariates, and allows the use of separate thresholds for inferring dependence as opposed to independence, which can significantly improve the reliability of the algorithm, as shown in the next section.

6 EMPIRICAL RESULTS

In this section we evaluate the performance of our approach as described in Sections 3 and 4, and empirically compare it with the simple approaches mentioned in Section 2, as well as to the method of VanderWeele and Shpitser (2011) (providing the necessary information about causes of x and y taken from the generating graph) and to the approach based on the FCI algorithm, as discussed in Section 5.[3] Matlab code to reproduce all results is available at `http://www.cs.helsinki.fi/u/entner/CovariateSelection/`.

We use linear Gaussian models throughout, testing for zero partial correlation to infer conditional independence, using Fisher's Z transformation. We first randomly create acyclic connections among the variables, satisfying the partial order assumption. Next we assign connection strengths to the edges and standard deviations to the error terms. We then generate data for each variable $v_i \in \mathcal{V}$ using $v_i = \sum_{v_j \in pa_i} b_{ij} v_j + e_i$, where pa_i denotes the parent set of v_i, b_{ij} the direct causal effect of v_j on v_i, and e_i the error term of v_i. Finally we hide the data over the latent variables in $\mathcal{U}$.

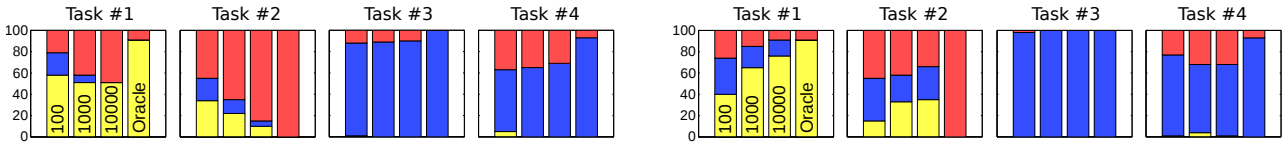
Table 1: Overview of the four simulation tasks.

Task	Effect	Admis. set	Inferences
#1	non-zero	yes	'±' or '?'
#2	non-zero	no	'?'
#3	zero	yes	'0'
#4	zero	no	'0' or '?'

We first use models with $|\mathcal{W}| = 10$ observed and $|\mathcal{U}| = 5$ hidden covariates, in which case it is possible to go through all $\mathcal{Z} \subseteq \mathcal{W}$, and all combinations of $w \in \mathcal{W}$ and $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$, and count the times rules R1 and R2 apply. We divide our simulations into four cases, summarized in Table 1. For task #1, there exists a non-zero effect of x on y, and there exists an admissible set (based on d-separation in the underlying graph). However, it may not be possible from the dependencies and independencies alone to conclude that there is a causal effect. Hence, for this task, while '±' is the appropriate conclusion in some instances, in others the appropriate inference is '?'. For task #2, there exists a causal effect but no admissible set, and in this case the only valid inference is '?'. In task #3, there is no causal effect but there exists an admissible set. In this case, any admissible set satisfies condition (i) of rule R2, so the only appropriate inference is '0'. Finally, in task #4, there is no causal effect and no admissible set. In some instances, conditions (ii) and (iii) of R2 nevertheless apply, so '0' can be inferred. In other instances, none of the conditions apply so the only valid inference is '?'. For each task, we randomly generate 100 models as described above.

We first compare the results of our approach to the results of the FCI based algorithm including the background knowledge given by the partial ordering (we actually use the CFCI algorithm, a conservative version of FCI, to obtain more reliable results). In Figure 2, we show the distributions of the obtained inferences for the four tasks with three different sample sizes and an independence oracle ('infinite sample size'). The results for our inference rules, presented in **(a)**, show that the larger the sample size gets, the fewer mistakes we make. In particular, in task #1, the erroneous inferences of '0' (in blue) decrease to an insignificant level as the sample size grows. For task #2, any inferences of '±' (in yellow), or '0' (in blue) are mistakes, and both these proportions decrease with growing sample size. Lastly, for tasks #3 and #4, there

[3] We do not compare to the method of de Luna et al. (2011), as their procedure solely aims at selecting a *minimal* set of covariates among a set which is known to be admissible. For the same reason, in the approach of VanderWeele and Shpitser (2011) we simply use the full set of covariates suggested by their selection procedure without trying to find a smaller subset.



(a) Inference rules

(b) FCI based approach

Figure 2: Inferences produced for each of the four tasks of Table 1 using 100 models with 10 observed and 5 hidden covariates. The first three bars in each plot display the inferences with 100, 1000, and 10000 samples, respectively, the last bar for an independence oracle ('infinite sample size'). The colored proportion of the bars represent how often the corresponding decision was made, with '±' in yellow, '0' in blue, and '?' in red.

are essentially no errors (no inference of '±'), and the number of inferences of '0' increases with sample size (though slowly). Compared to our rules, from **(b)** we see that the FCI based approach performs as well (or better) on tasks #3 and #4. However, it makes a significant amount of mistakes (inference of '0', in blue) on task #1, even for large sample sizes. On task #2, where the only correct inferences are '?', the number of errors is large, and growing with sample size.

As the main objective is to obtain accurate estimates of the causal effect, we give in Figure 3 boxplots of the difference between the true effect and the estimated effect for the various approaches. For task #1, where there truly exists an admissible set, the method of VanderWeele and Shpitser (2011) always yields an admissible set, and so tends to be the most reliable method; however, most of the other approaches are almost as good. For task #2 on the other hand, the benefit of our rules is clear: With the largest sample size, only in 15 out of 100 instances did the approach make a prediction, and for those estimates the errors tend to be smaller on average than for the other methods. Since it is impossible to know, a priori, to which task a given problem belongs, it is crucial to avoid making predictions when the data does not warrant making one. In tasks #3 and #4, both our approach and that based on FCI are essentially avoiding any errors. (Note that these two approaches actually set the estimated causal effect to 0, while the other approaches simply estimate the regression coefficient with adjustment; this explains some of the discrepancy.) The main advantage is again that our approach avoids making a prediction when one is not warranted by the data, giving fewer but much more accurate predictions.

We have also tested our approach on models with 100 observed covariates, using the sampling approach described in Section 4. The main result is that, while the absence of a causal effect (tasks #3 and #4) can be detected reliably, and any non-zero estimated effects in tasks #1 and #2 have little error, there are many cases in tasks #1 and #2 in which our approach erroneously infers '0'. A figure and more details are provided in the Supplementary Material.

7 CONCLUSIONS

Finding an appropriate adjustment set among the covariates $\mathcal{W}$ to obtain a consistent and unbiased estimator of the 'treatment' x on the 'outcome' y, from non-experimental data, is an important problem. For acyclic models with no selection bias, and when the underlying graph is known, this problem is solved. We have considered the difficult case where the causal structure is unknown. Assuming acyclic connections among the variables, faithfulness, no selection bias, and that all observed covariates are pre-treatment, we have presented two simple rules to infer whether there is a non-zero effect of x on y (and an admissible set for adjustment), a zero effect of x on y, or whether this information is impossible to determine from the data alone. We have shown these inference rules to be sound and complete, and demonstrated in simulations their advantage compared to other proposed approaches.

Many questions are still left unanswered. For the practical implementation of the algorithm, several shortcomings were already mentioned in Section 4. Improving the ad-hoc procedure of using a Bayesian classifier for propagating uncertainty would be an interesting avenue of future research. It would also be important to verify the performance of the method when applied to data that is not linear-Gaussian.

Another open problem is how to incorporate further background knowledge, such as that of VanderWeele and Shpitser (2011) (having information on which variables are causes of x or y), or of Spirtes and Cooper (1999), and Chen et al. (2007) (knowing that certain variables are exogenous). One can for instance prove the following theorem, which allows inferring that certain sets are *not* admissible. (Proof in the Supplementary Material.)

Theorem 4. ***(Test for non-admissibility)*** *Given the model and partial ordering assumptions of Section 2, if there exists an* exogenous *variable* $w \in \mathcal{W}$ *and a (possibly empty) set* $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ *such that*

(i) $w \not\perp\!\!\!\perp x \mid \mathcal{Z}$, *and*
(ii) $w \not\perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

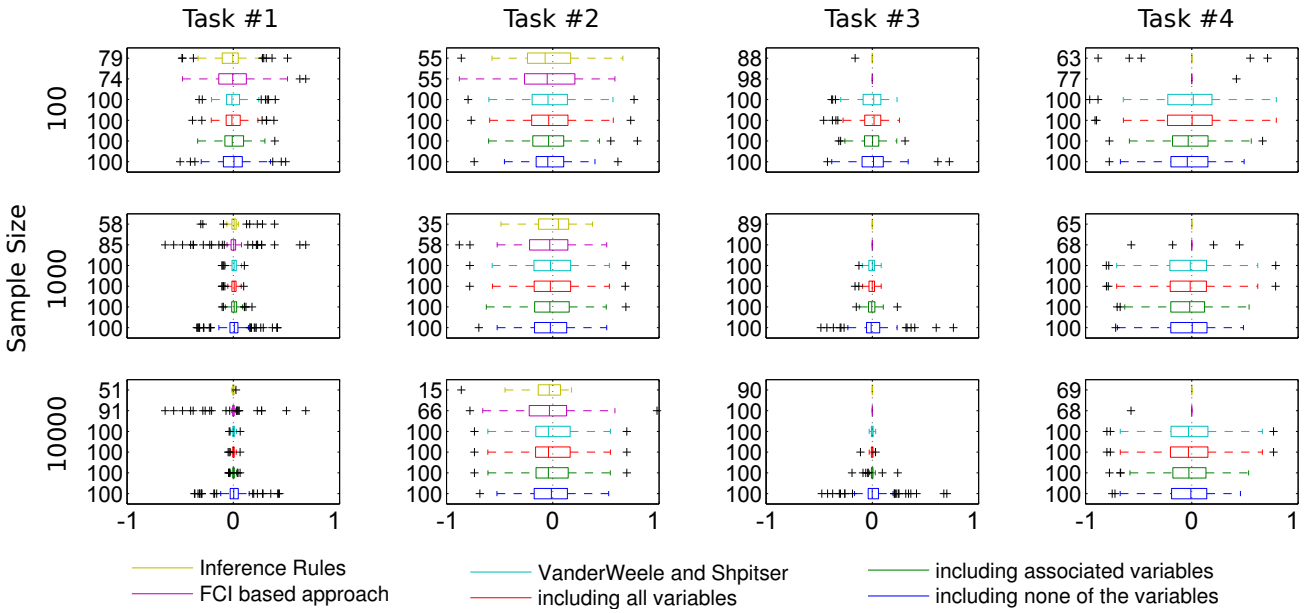


Figure 3: Boxplots of the differences between the true and estimated causal effect, shown along the horizontal axis, using 100 models with 10 observed and 5 hidden covariates. In the rows are plots for various sample sizes (100, 1000, 10000), in the columns for the four tasks of Table 1. In each subfigure we show the boxplots for the methods as indicated in the legend. For the boxplots, the box ranges from q_1, the 1^{st} quartile, to q_3, the 3^{rd} quartile, and the median is indicated by a horizontal line in the box. The whiskers extend to the furthest point not considered as an outlier. Outliers are all points larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$. The numbers along the vertical axis indicate how often the corresponding method output an estimate.

then the set $\mathcal{Z}$ *is* not *admissible.*

Furthermore, one could relax the assumption of knowing a partial ordering among the observed variables. The most straightforward generalization would be to drop the requirement that y cannot be a causal ancestor of x, but keep the knowledge that the covariates $\mathcal{W}$ causally precede both x and y, as in the work of Spirtes and Cooper (1999), and Chen et al. (2007). A second way to relax this assumption would be to allow the covariates $\mathcal{W}$ to be between x and y, keeping the knowledge that x precedes y. This could lead to opportunities to identify additional causal effects, as well as to distinguish between direct and indirect effects.

Lastly, we want to mention that especially in non-linear or discrete data sets, the size of the admissible set $\mathcal{Z}$ can significantly influence the performance of the estimator, and small sets $\mathcal{Z}$ are desirable. Thus, combining our approach with further selection criteria to obtain a smaller or minimal admissible set, such as discussed in de Luna et al. (2011), or VanderWeele and Shpitser (2011), would be very useful.

Acknowledgments

DE and POH were supported by the Academy of Finland.

APPENDIX: PROOFS

Proof of Theorem 2 (Soundness)

We need to show that whenever either rule R1 or R2 applies, then the resulting inference is correct.

R1: We will show that, if rule R1 applies, then there is a non-zero effect of x on y, and $\mathcal{Z}$ blocks all back-door paths from x to y, and hence the set $\mathcal{Z}$ is admissible.

Condition (i) ensures that there exist active paths from w to y given $\mathcal{Z}$, and condition (ii) ensures that all such paths must pass through x (as a non-collider, and by assumption into x), since otherwise adding x to the conditioning set would not block all the paths.

Since by assumption x is not an ancestor of any member of $\mathcal{W}$, and in condition (i) x is not in the conditioning set, any active path from w to y given $\mathcal{Z}$ must include the directed edge from x to y, and hence, by faithfulness, the effect of x on y is non-zero.

Together, conditions (i) and (ii) thus imply that there exists at least one active path from w to x given $\mathcal{Z}$, pointing into x. Now, if there existed an active back-door path from x to y given $\mathcal{Z}$, then concatenating these two paths would yield an active path from w to y given $\mathcal{Z} \cup \{x\}$ (using Lemma 3.3.1 of Spirtes et al. (2000) if the paths have more than one node in com-

mon), since there is a collider at x. This would violate condition (ii), and hence, all back-door paths from x to y must be blocked by $\mathcal{Z}$.

R2: Assume that the causal effect of x on y is not zero, i.e. the model contains the edge $x \to y$. Faithfulness would then ensure that this yields a dependence between x and y given *any* set $\mathcal{Z}$. Thus, in this case condition (i) would never hold.

Condition (ii) ensures that there is an active path π from w to x, given $\mathcal{Z}$, pointing into x, since by assumption $y \notin \mathcal{Z}$ so all paths via y are blocked by colliding arrows. Now, if there was an arrow from x to y then appending this arrow to π would lead to an active path $w \ast\!\!\to \ldots \ast\!\!\to x \to y$, given $\mathcal{Z}$, since $x \notin \mathcal{Z}$. But this is in contradiction to condition (iii). Thus, there cannot be an arrow from x to y. □

Proof of Theorem 3 (Completeness)

We show that whenever neither rule R1 nor R2 applies one cannot, only based on the conditional independencies and dependencies among the observed variables, know whether x has a causal effect on y or not.

'±': First consider the case where x really is a cause of y, i.e. the true model contains the edge $x \to y$. We will examine two separate cases:

(a) Assume that there does *not* exist a confounder $u \in \mathcal{U}$ that is a parent of both x and y. Consider adding such a confounder to the model. We will show that if such an addition changes any (conditional) independencies or dependencies among the observed variables $\mathcal{W} \cup \{x, y\}$, then rule R1 applies.

First note that adding a confounder of the above form can only change a conditional independence to a dependence, and not the reverse. Also note that any dependence created by the added confounder must rely on unblocked paths that all contain the subpath $w \ast\!\!\to x \leftarrow u \to y$, with $w \in \mathcal{W}$. To result in a change from an independence to a dependence, conditional on some set $\mathcal{Z} \cup \{x\}$, there can be no previously active path from w to y given $\mathcal{Z} \cup \{x\}$. Hence condition (ii) of R1 must apply. Since by necessity there is also an active (sub)path from w to x given $\mathcal{Z}$, and by the fact that the model contains the edge $x \to y$ condition (i) holds, so rule R1 applies.

(b) Next, assume that there *does* exist a confounder $u \in \mathcal{U}$ that is a parent of both x and y. We will now show that it is possible to remove the edge $x \to y$, and add compensating edges to the model, such that all (conditional) independencies and dependencies among the observed variables remain unaltered.

Consider the model obtained by removing the edge $x \to y$ and adding edges $v \to y$ from all $v \in \mathcal{V}$ that are parents of x in the true graph (note that some of these v may be unobserved). Any active path broken by removing the edge $x \to y$ must include a subpath $v \to x \to y$, with v a parent of x, and x not in the conditioning set. All such previously active paths are restored by adding the edge $v \to y$. Furthermore, adding the edge $v \to y$ can never create any new active paths, because v and y were always d-connected in the original model because it included both paths $v \to x \to y$ and $v \to x \leftarrow u \to y$. Hence all independencies and dependencies remain the same.

Together, parts (a) and (b) above show that for *any* true model containing the edge $x \to y$, either (i) it is possible to find (by if necessary combining the model alterations in (a) and (b)) an alternative model which does *not* contain the edge $x \to y$ yet yields the same set of independencies and dependencies, or (ii) rule R1 applies. This implies that in all cases where x is truly a cause of y, we can either detect it (and provide an admissible set) or it is undecidable from the data whether x is a cause of y or not.

'0': Next, consider the case where x is *not* a cause of y, i.e. the true model does not contain the edge $x \to y$. We will show that if adding this edge changes any independencies or dependencies, then rule R2 applies. Thus, if R2 does not apply, it cannot be known from independencies and dependencies whether x is a cause of y or not.

Adding the edge $x \to y$ can only turn independencies into dependencies, not the reverse. First consider independencies of the form $x \perp\!\!\!\perp y \,|\, \mathcal{Z}$, with $\mathcal{Z} \subseteq \mathcal{W}$, which are obviously turned into dependencies when the edge is added. Any such cases are handled by condition (i) of rule R2. Now, if x and y are dependent conditioned on all subsets of $\mathcal{W}$ so that condition (i) does not apply, then no independence between x and any $w \in \mathcal{W}$ can be affected by the addition of the edge $x \to y$, because an active path with an arrowhead at y already existed prior to the addition. Hence we only need to consider independencies between any $w \in \mathcal{W}$ and y. If x is in the conditioning set then the addition of the edge $x \to y$ cannot change any such independencies. Thus, the only new dependencies created must rely on active paths that must contain the subpath $w \ast\!\!\to x \to y$, and x must not be in the conditioning set. This corresponds to the combination of conditions (ii) and (iii) of rule R2. Hence if any new dependencies would be created by adding the edge $x \to y$, then R2 applies.

In summary, we have shown that whether or not the true model contains the edge $x \to y$ or not, if neither rule R1 nor R2 applies, one cannot based on conditional independence and dependence information reliably detect whether the edge is present or not. □

References

Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007). Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8(R219).

de Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Entner, D., Hoyer, P. O., and Spirtes, P. (2012). Statistical test for consistent estimation of causal effects in linear non-gaussian models. In *JMLR Workshop and Conference Proceedings 22 (AISTATS-2012)*.

Pearl, J. (2009). *Causality: Models, Reasoning, and inference*. Cambridge University Press, 2nd edition.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. In *22nd Conference on Uncertainty in Artificial Intelligence*.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *26th Conference on Uncertainty in Artificial Intelligence*.

Spirtes, P. and Cooper, G. F. (1999). An experiment in causal discovery using a pneumonia database. In *International conference on Artificial Intelligence and Statistics*.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge MA: MIT Press, 2nd edition.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67:1406–1413.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896.