Causal Inference on Discrete Data using Additive Noise Models

Jonas Peters, Dominik Janzing and Bernhard Schölkopf

***Abstract*— Inferring the causal structure of a set of random variables from a finite sample of the joint distribution is an important problem in science. The case of two random variables is particularly challenging since no (conditional) independences can be exploited. Recent methods that are based on additive noise models suggest the following principle: Whenever the joint distribution $\mathbf{P}^{(X,Y)}$ admits such a model in one direction, e.g. $Y = f(X) + N,\ N \perp\!\!\!\perp X$, but does not admit the reversed model $X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y$, one infers the former direction to be causal (i.e. $X \to Y$). Up to now these approaches only deal with continuous variables. In many situations, however, the variables of interest are discrete or even have only finitely many states. In this work we extend the notion of additive noise models to these cases. We prove that it almost never occurs that additive noise models can be fit in both directions. We further propose an efficient algorithm that is able to perform this way of causal inference on finite samples of discrete variables. We show that the algorithm works both on synthetic and real data sets.**

***Index Terms*— Causal Inference, Regression, Graphical Models**

I. Introduction

Inferring causal relations between random variables from observed data is a challenging task if no controlled randomized experiments are available. So-called constraint-based approaches to causal discovery (Pearl, 2000; Spirtes et al., 2000) select among all directed acyclic graphs (DAGs) those that satisfy the Markov condition and the faithfulness assumption. These conditions relate the graph structure to the observed distribution: Roughly speaking, the graph is *Markov* if all (conditional) independences imposed by the graph structure can be found in the distribution and *faithful* if all (conditional) independences that can be found in the distribution are imposed by the graph structure. Those constraint-based approaches are unable to distinguish among causal DAGs that impose the same independences (Markov equivalence classes, Verma and Pearl (1991)). In particular, it is impossible to distinguish between $X \to Y$ and $Y \to X$.

More recently, several methods have been suggested that do not only use conditional independences, but also more sophisticated properties of the joint distribution. We explain these ideas for the two variable setting. Shimizu et al. (2006); Kano & Shimizu (2003) use models

$$Y = f(X) + N\,, \tag{1}$$

where f is a linear function and N is additive noise that is independent of the hypothetical cause X. This is an example for an additive noise model (ANM) from X to Y. Apart from trivial cases, $\mathbf{P}^{(X,Y)}$ can only admit such a model from X to Y and from Y to X in the bivariate Gaussian case. We say the model is identifiable in the "generic case". (In the remainder of the article we will use "genericness" in the meaning of "there are almost no exceptions"; for the precise statement we refer to the cited literature.) They propose the following inference principle to distinguish between cause and effect: Whenever such an ANM exists in one direction but not in the other, one infers the former to be the causal direction.

Janzing & Steudel (2010) give theoretical support for this principle using the concept of Kolmogorov complexity. Peters et al. (2009) apply the concept of ANMs to ARMA time series in order to detect whether a sample of a time series has been reversed. Hoyer et al. (2009); Mooij et al. (2009) generalize the method to non-linear functions f and showed that generic models of this form generate joint distributions that do not admit such an ANM from Y to X (here, genericness means that the triple f and the densities of X and noise N do not satisfy a very specific differential equation). Zhang & Hyvarinen (2009) augment the model by applying an invertible non-linear function g to the right-hand side of equation (1) and still obtain identifiability in the generic case. Janzing et al. (2009) make first steps towards identifying hidden common causes. All these proposals, however, were only designed for continuous variables X and Y.

For discrete variables, Sun et al. (2008) propose a method to measure the complexity of causal models via a Hilbert space norm of the logarithm of conditional densities and prefer models that induce smaller norms. Sun et al. (2006) fit joint distributions of cause and effect with conditional densities whose logarithm is a second order polynomial (up to the log-partition function) and show that this often makes causal directions identifiable when some or all variables are discrete. For discrete variables, several Bayesian approaches (Heckerman et al., 1999) are also applicable, but the construction of good priors are challenging and often the latter are designed such that Markov equivalent DAGs still remain indistinguishable.

Here, we extend the model in equation (1) to the discrete case in two different ways: (A) If X and Y take values in $\mathbb{Z}$ (the support may be finite, though) ANMs can be defined analogously to the continuous case. (B) If X and Y take only finitely many values we can also define ANMs by interpreting the $+$ sign as an addition in the finite ring $\mathbb{Z}/m\mathbb{Z}$. We propose to apply this method to variables where the cyclic structure is appropriate (e.g., the direction of the wind after discretization, day of the year, season). Remark 1 in section II-B describes how the second model can also be applied to structureless sets; this may be helpful whenever the random variables are categorical and when these categories do not inherit any kind of ordering (e.g. different treatments of organisms or phenotypes). In the following article we refer to (A) by *integer models* and to (B) by *cyclic models*.

We adopt the causal inference method from above: If there is an ANM from X to Y, but not vice versa, we propose that X is causing Y (more details in section II). Such a procedure is

All authors are affiliated to MPI for Biological Cybernetics, Tübingen, Germany

sensible if there are only few instances, in which there are ANMs in both directions. If, for example, all ANMs from X to Y also allow for an ANM from Y to X, we could not draw any causal conclusions at all. In section III we show that these *reversible* cases are very rare and thereby answer this theoretical question.

For a practical causal inference method we have to test whether the data admit an ANM. We propose an efficient procedure that proved to work well in practice (section IV).

Note that a shortened version of this work has already been published by Peters et al. (2010). In addition, here we cover the "cyclic case" (denoted above by B), provide proofs and more experiments, investigate the binary case separately, analyze the run-time of the algorithm empirically and give an outlook to generalizations of discrete ANMs.

The paper is organized as follows: In section II we extend the concept of ANMs to discrete random variables and show the corresponding identifiability results in section III. In section IV we introduce an efficient algorithm for causal inference on finite data, for which we show experimental results in section V. Section VI contains the proofs and section VII our conclusions.

II. Additive Noise Models for Discrete Variables

As it has been proposed for the continuous case by Shimizu et al. (2006); Hoyer et al. (2009); Zhang & Hyvarinen (2009) we assume the following causal principle to hold throughout the remainder of this article:

Causal Inference Principle (for discrete random variables) *Whenever Y satisfies an additive noise model with respect to X and not vice versa then we infer X to be the cause for Y, and we write $X \to Y$.*

Note that whenever there is no additive noise model in any direction (which may well happen) the method remains inconclusive and other causal inference methods should be tried.

There are two reasons why we do not expect the true data generating process to allow an ANM *only* in the wrong causal direction: (1) We hope that nature prefers "simple" mechanisms (Occam's Razor). (2) Janzing & Steudel (2010) use the concept of Kolmogorov complexity to show that this can only be the case if the cause distribution $p(\mathrm{cause})$ and the mechanism $p(\mathrm{effect}|\mathrm{cause})$ are matched in a precise way, whereas we rather expect input and mechanism to be most often "independent" (although there might exist cases, for which this assumption is violated).

Now we precisely explain what we mean by an additive noise model in the case of discrete random variables. For simplicity we denote $p(x) = \mathbf{P}(X = x)$, $q(y) = \mathbf{P}(Y = y)$, $n(l) = \mathbf{P}(N = l)$ and $\tilde{n}(k) = \mathbf{P}(\tilde{N} = k)$ and $\operatorname{supp} X$ is defined as $\operatorname{supp} X := \{k \,|\, p(k) > 0\}$.

A. Integer Models

Assume that X and Y are two random variables taking values in $\mathbb{Z}$ (their distributions may have finite support). We say that there is an additive noise model (ANM) from X to Y if there is a function $f : \mathbb{Z} \to \mathbb{Z}$ and a noise variable N such that the joint distribution $\mathbf{P}^{(X,Y)}$ allows to write

$$Y = f(X) + N \quad \text{and } N \perp\!\!\!\perp X\,.$$

Furthermore we require $n(0) \geq n(j)$ for all $j \neq 0$. This does not restrict the model class, but is due to a freedom we have in choosing f and N: If $Y = f(X)+N$, $N \perp\!\!\!\perp X$, then we can always construct a new function f_j, such that $Y = f_j(X)+N_j$, $N_j \perp\!\!\!\perp X$ by choosing $f_j(i) = f(i) + j$ and $n_j(i) = n(i + j)$.
Such an ANM is called *reversible* if there is also an ANM from Y to X, i.e. if it satisfies ANMs in both directions.

B. Cyclic Models

We can extend ANMs to random variables which inherit a cyclic structure and therefore take values in a periodic domain. Random variables are usually defined as measurable maps from a probability space into the real numbers. Thus, we first make the following definition

Definition 1: Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A function $X : \Omega \to \mathbf{Z}/m\mathbf{Z}$ is called an m-cyclic random variable if $X^{-1}(k) \in \mathcal{F}\ \forall k \in \mathbf{Z}/m\mathbf{Z}$. All other concepts of probability theory (like distributions and expectations) can be constructed analogously to the well-known case, in which X takes values in $\{0, \ldots, m-1\}$.

Let X and Y be m- and $\tilde{m}$-cyclic random variables, respectively. We say that Y satisfies an ANM from X to Y if there is a function $f : \mathbf{Z}/m\mathbf{Z} \to \mathbf{Z}/\tilde{m}\mathbf{Z}$ and an $\tilde{m}$-cyclic noise N such that

$$Y = f(X) + N \quad \text{and } N \perp\!\!\!\perp X\,.$$

Again we require $n(0) \geq n(j)$ for all $j \neq 0$ and call this model *reversible* if there is a function $g : \mathbf{Z}/\tilde{m}\mathbf{Z} \to \mathbf{Z}/m\mathbf{Z}$ and an m-cyclic noise $\tilde{N}$ such that $X = g(Y) + \tilde{N}$ and $\tilde{N} \perp\!\!\!\perp Y$.

Remark 1: Cyclic models are not restricted to random variables that take integers as values: Assume that X and Y take values in $\mathcal{A} := \{a_1, \ldots, a_m\}$ and $\mathcal{B} := \{b_1, \ldots, b_{\tilde{m}}\}$, which are structureless sets. Considering functions $f : \mathcal{A} \to \mathcal{B}$ and models with $\mathbf{P}(Y = b_j \,|\, X = a_i) = p$ if $b_j = f(a_i)$ and $(1-p)/(\tilde{m}-1)$ otherwise, is a special case of an ANM: Impose any cyclic structure on the data and use the additive noise $\mathbf{P}(N = 0) = p, \mathbf{P}(N = l) = (1-p)/(\tilde{m}-1)$ for $l \neq 0$.

C. Relations

The following two remarks are essential in order to understand the relationship between integer and cyclic models: (1) The difference between these two models manifests in the target domain. If we consider an ANM from X to Y it is important whether we put integer or cyclic constraints on Y (and thus on N). It does not make a difference, however, whether we consider the regressor X to be cyclic (with a cycle larger than $\#\operatorname{supp} X$) or not. The independence constraint remains the same. (2) In the finite case ANMs with cyclic constraints are more general than integer models: Assume there is an ANM $Y = f(X) + N$, where all variables are taken to be non-cyclic and Y takes values between k and l, say. Then we still have an ANM $Y = f(X)+N$ if we regard Y to be $l-k+1$-cyclic because $N \mod (l-k+1)$ remains independent of X. It is possible, however, that $N \not\perp\!\!\!\perp X$, but $N \mod (l-k+1) \perp\!\!\!\perp X$ (as shown in Example 2).

III. Identifiability

Whether or not there is an ANM between X and Y only depends on the form of the joint distribution $\mathbf{P}^{(X,Y)}$. Let A be the set of all possible joint distributions and F its subset that allows an additive noise model from X to Y in the "forward direction", whereas B allows an ANM in the backward direction from Y to X (see Figure 1).

Some trivial examples like $p(0) = 1, n(0) = 1$ and $f(0) = 0$ immediately show that there are joint distributions allowing ANMs in both directions, meaning $F \cap B \neq \emptyset$. But how large is this intersection? The proposed method would not be useful if we find out that F and B are almost the same sets. Then in most cases ANMs can be fit either in both directions or in none. Both, for ANMs with integer constraints and with cyclic constraints we identify the intersection $F \cap B$ and show that it is indeed a very small set. Imagine, we observe data from a natural process that allows an ANM in the causal direction. If we are "unlucky" and the data generating process happens to be in $F \cap B$, our method does not give wrong results, but answers "I do not know the answer".

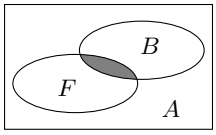


Fig. 1. How large is $F \cap B$?

A. Integer Models

1) Y or X has finite support: First we assume that either the support of X or the support of Y is finite. This already covers most applications. Figure 2 (the dots indicate a probability greater than 0) shows an example of a joint distribution that allows an ANM from X to Y, but not from Y to X. This can be seen easily at the "corners" $X = 1$ and $X = 7$: Whatever we choose for $g(0)$ and $g(4)$, the distribution of $\tilde{N} \,|\, Y = 0$ is supported only by one point, whereas $\tilde{N} \,|\, Y = 4$ is supported by 3 points. Thus $\tilde{N}$ cannot be independent of Y. Figure 3 shows a (rather non-generic) example that allows an ANM in both directions if we choose $p(a_i) = \frac{1}{36}, p(b_i) = \frac{2}{36}$ for $i = 1, \ldots, 4$ and $p(a_i) = \frac{2}{36}, p(b_i) = \frac{4}{36}$ for $i = 5, \ldots, 8$. We prove the following

Theorem 1: Assume either X or Y has finite support. An ANM $X \to Y$ is reversible $\iff$ there exists a disjoint decomposition $\bigcup_{i=0}^{l} C_i = \operatorname{supp} X$, such that a) - c) are satisfied:

a) The C_is are shifted versions of each other

$$\forall i \, \exists d_i \geq 0 \, : \, C_i = C_0 + d_i$$

and f is piecewise constant: $f \,|_{C_i} \equiv c_i \; \forall i$.

b) The probability distributions on the C_is are shifted and scaled versions of each other with the same shift constant as above: For $x \in C_i$, $\mathbf{P}(X = x)$ satisfies

$$\mathbf{P}(X = x) = \mathbf{P}(X = x - d_i) \cdot \frac{\mathbf{P}(X \in C_i)}{\mathbf{P}(X \in C_0)}.$$

c) The sets $c_i + \operatorname{supp} N := \{c_i + h \, : \, n(h) > 0\}$ are disjoint.

(Note that such a decomposition satisfying the same criteria also exists for $\operatorname{supp} Y$ by symmetry.) In the example of Figure 3 all a_i belong to C_0, all b_j to C_1 and $d_1 = 1$. As for the other theorems of this section the proof is provided in section VI. Its main point is based on the asymmetric effects of the "corners" of the joint distribution. In order to allow for an infinite support of X (or Y) we will thus generalize this concept of "corners".

Theorem 1 provides a full characterization of cases that allow for an ANM in both directions. Each of the conditions is very restrictive by itself, all conditions together describe a very small class of models: in almost all cases the direction of the model is identifiable. We have the following corollary:

Corollary 2: Consider a discrete ANM from X, which takes values $x_1, \ldots, x_m$ ($m > 1$), to Y with a non-constant function f (otherwise X and Y are independent). Let the noise N take values from $N_{\min}$ to $N_{\max}$ and put any prior measure on the parameters $n(k)$ for $k = N_{\min}, \ldots, N_{\max}$ and $p(x_k), k = 1, \ldots, m$ that is absolutely continuous to the Lebesgue measure. If further $\min_{i,j \in \{1,\ldots,m\} \,:\, i \neq j} f(x_i) - f(x_j) \leq N_{\max} - N_{\min}$ we have the following statement: Only a parameter set of measure 0 admits an ANM from Y to X.

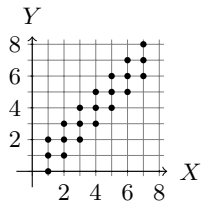


Fig. 2. This joint distribution satisfies an ANM only from X to Y.

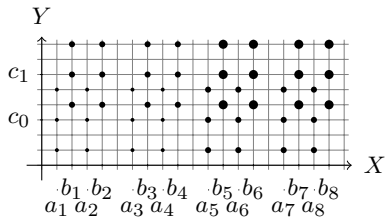


Fig. 3. Only carefully chosen parameters allow ANMs in both directions. (Radii correspond to probability values.)

2) X and Y have infinite support:

Theorem 3: Consider an ANM $X \to Y$ where both X and Y have infinite support. We distinguish between two cases

a) **N has compact support:** $\exists m, l \in \mathbb{Z}$, s.t. $\operatorname{supp} N = [m, l]$. Assume there is an ANM from X to Y and f does not have infinitely many infinite sets, on which it is constant. Then we have the following equivalence: The model is reversible if and only if there exists a disjoint decomposition $\bigcup_{i=0}^{\infty} C_i = \operatorname{supp} X$ that satisfies the same conditions as in Theorem 1.

b) **N has entire $\mathbb{Z}$ as support:** $\mathbf{P}(N = k) > 0 \, \forall \, k \in \mathbb{Z}$. Suppose X and Y are dependent and there is a reversible ANM $X \to Y$. Fix any $m \in \mathbb{Z}$. If f, $\mathbf{P}^N$ and $p(k)$ for all $k \geq m$ are known, then all other values $p(k)$ for $k < m$ are determined. That means even a small fraction of the parameters determine the remaining parameters.

Note that the first case is again a complete characterization of all instances of a joint distribution, an ANM in both directions is conform with. The second case does not yield a complete characterization, but shows how restricted the choice of a distribution $\mathbf{P}^X$ is (given f and $\mathbf{P}^N$) that yields a reversible ANM.

B. Cyclic Models

Assume $Y = f(X) + N$ with $N \perp\!\!\!\perp X$. We will show that in the generic case the model is still not reversible, meaning there is no g and $\tilde{N}$, such that $X = g(Y) + \tilde{N}$ with $\tilde{N} \perp\!\!\!\perp Y$. However, as mentioned in section II-C, in finite domains this model class is larger than the class of integer models. We will see that correspondingly also the number of reversible cases increases.

Note that the model $Y = f(X) + N$ is reversible if and only if there is a function g, such that

$$p(x) \cdot n\big(y - f(x)\big) = q(y) \cdot \tilde{n}\big(x - g(y)\big) \; \forall x, y \,, \qquad (2)$$

where $q(y) = \sum_{\tilde{x}} p(\tilde{x}) n(y - f(\tilde{x}))$ and $\tilde{n}(a) = p(g(\tilde{y}) + a) \cdot n(\tilde{y} - f(g(\tilde{y}) + a))/q(\tilde{y}) \quad \forall \tilde{y} \, : \, q(\tilde{y}) \neq 0$.

1) Non-Identifiable Cases: First, we give three (characteristic) examples of ANMs that are not identifiable. This restricts the class of situations in which identifiability can be expected. Figure 4 shows instances of Examples 1 and 2.

Example 1: Independent X and Y always admit an ANM from X to Y and from Y to X. We therefore have:

(i) If $Y = f(X) + N$ and $f(k) = \text{const}$ for all $k : p(k) \neq 0$, then the model is reversible.

(ii) If $Y = f(X) + N$ for a uniformly distributed noise N, then the model is reversible.

Proof: In both cases it X and Y are independent. Thus, $X = g(Y) + X$ with $g \equiv 0$ is a backward model. ■

Example 2: If $Y = f(X) + N$ for a bijective and affine f and uniformly distributed X, then the model is reversible.

Proof: Since X is uniform and $f(x) = ax + b$ is bijective, Y is uniform, too. For $g(y) = f^{-1}(y)$ and $\tilde{n}(k) = n(b - f(k)) = n(y - f(g(y) + k))$ equation (2) is satisfied. ■

Example 3: We give two more examples of non-identifiable cases that show why an if-and-only-if characterization as in Theorem 1 is hard to obtain:

(i) Figure 5 (left) shows an example, where the sets on which f is constant neither satisfy condition c) nor are they shifted versions of each other.

(ii) The same holds for Figure 5 (right), this time even satisfying the additional constraint that $\mathbf{P}(N = 0) > \mathbf{P}(N = k) \forall k \neq 0$. Here, X is not uniformly distributed, either.

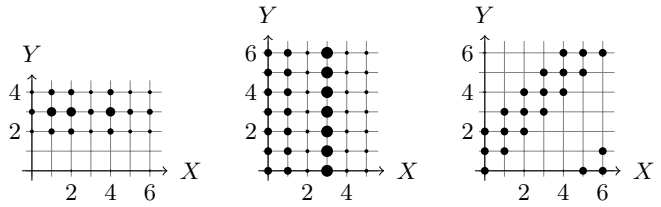


Fig. 4. These joint distributions allow ANMs in both directions. They are instances of Examples $1(i)$, $1(ii)$ and 2 (from left to right).

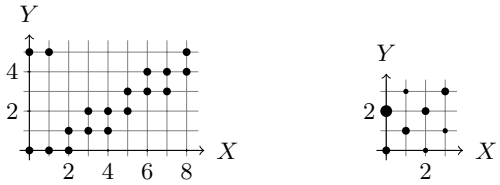


Fig. 5. These joint distributions allow ANMs in both directions. They are instances of Examples 3 (i) (left) and (ii) (right).

2) Identifiability Results: The counter examples from above already show that cyclic models are in some aspect more difficult than integer models and we thus do not provide a full characterization of all reversible cases as we have done in the integer case. Nevertheless, we provide necessary conditions for reversibility, which is sufficient for our purpose.

Usually the distribution $n(l)$ (similar for $p(k)$) is determined by $\tilde{m} - 1$ free parameters. As long as the sum remains smaller than 1, there are no (equality) constraints for the values of $n(0), \ldots, n(\tilde{m}-2)$. Only $n(\tilde{m}-1)$ is determined by $\sum_{l=0}^{\tilde{m}-1} n(l) = 1$. We show that in the case of a reversible ANM the number of free parameters of the marginal $n(l)$ is heavily reduced. The exact number of constraints depends on the possible backward functions g, but can be bounded from below by 2. Furthermore the proof shows that a "dependence" between values of p and n is introduced. Both of these constraints are considered to lead to non-generic models. That means for any *generic* choice of p and n we can only have an ANM in one direction.

Note further that $(\#\text{supp}\, X \cdot \#\text{supp}\, N)$ is the number of points (x, y) that have probability greater than 0. It must be possible to distribute these points equally to all points from $\#\text{supp}\, Y$ in order to allow a backward ANM. Thus we have the necessary condition $\#\text{supp}\, Y \,|\, (\#\text{supp}\, X \cdot \#\text{supp}\, N)$. (Here, $a \,|\, b$ denotes "a divides b", which we write if $\exists z \in \mathbb{Z} : b = z \cdot a$, and should not be confused with conditioning on a random variable.)

Theorem 4: Assume $Y = f(X) + N, N \perp\!\!\!\perp X$ with non-uniform X (m-cyclic), Y ($\tilde{m}$-cyclic) and N ($\tilde{m}$-cyclic) and non-constant f.

(i) There can only be an ANM from Y to X if $\#\text{supp}\, Y \,|\, (\#\text{supp}\, X \cdot \#\text{supp}\, N)$.

(ii) Assume that $\#\text{supp}\, X = m, \#\text{supp}\, N = \tilde{m}$. If there is an ANM from Y to X, at least one additional equality constraint is introduced to the choice of either p or n.

Again, the proof can be found in section VI.

C. Special Case: X and Y binary

We now investigate a special case, where X and Y are constrained to take binary values with probabilities $a := \mathbf{P}(X = 0, Y = 0), b := \mathbf{P}(X = 1, Y = 0), c := \mathbf{P}(X = 0, Y = 1)$ and $d := \mathbf{P}(X = 1, Y = 1)$. For this case we can compute a full characterization of reversible and irreversible ANMs. Therefore we assume the variables to be non-degenerate (i.e. $0 < \mathbf{P}(X = 0) = a + c < 1$ and $0 < \mathbf{P}(Y = 0) = a + b < 1$) and we use the following Lemma:

Lemma 5: Let N and X be non-degenerate binary variables. Then $N \perp\!\!\!\perp X \Leftrightarrow \mathbf{P}(N = 1 \,|\, X = 0) = \mathbf{P}(N = 1 \,|\, X = 1)$.

The integer model is not very informative. The only two possibilities to form an ANM with integer constraints is to choose deterministic noise or a constant function f. Clearly, both cases lead to reversible ANMs. More interestingly, the results for the cyclic case are non-trivial:

1) f is constant.
   Here, X and Y are independent and the ANM is thus reversible (see Example $1(i)$). Lemma 5 implies that $X \perp\!\!\!\perp N$ if and only if $\frac{c}{a+c} = \frac{d}{b+d}$. And this holds if and only if
   $$ad = bc$$
   (Here, neither of the parameters can be zero.)
2) f is non-constant.
   Without loss of generality let f be the identity function (we can always add an additive shift). This time we have $X \perp\!\!\!\perp N$ if and only if $\frac{c}{a+c} = \frac{b}{b+d}$, which is equivalent to
   $$ab = cd$$
   still assuming $a + c \neq 0 \neq b + d$.

Using symmetry it follows that there is an ANM from Y to X if and only if we have either $ac = bd$ or $ad = bc$.

We thus summarize (recall that only b and c or a and d can be zero at the same time):

- $ab = cd$ or $ad = bc$ leads to an ANM from X to Y.
- $ac = bd$ or $ad = bc$ leads to an ANM from Y to X.
- $a = d$ and $b = c$ (this implies uniform X and Y) or $a = d = 0$ or $b = c = 0$ or $ad = bc$ leads to a reversible ANM.

This also fits with the theoretical result of Proposition 6 in section VI: for bijective f and g (which is the only case that does not lead to independent X and Y) only uniformly distributed X

and Y lead to reversible ANMs. Using $d = 1 - a - b - c$ one can plot these conditions as surfaces (see Figures 6 and 7).

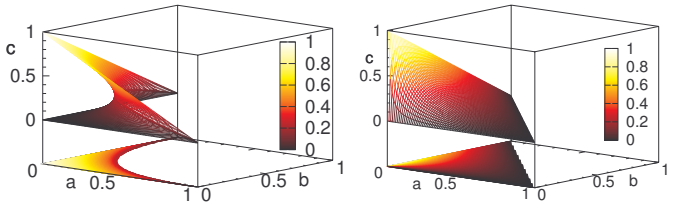


Fig. 6. For $X \not\perp\!\!\!\perp Y$ (both binary) these plots visualize the constraints of the joint distribution $\mathbf{P}^{(X,Y)}$ in order to allow for an ANM: either from X to Y ($ab = cd$, left) or from Y to X ($ac = bd$, right). Note that the both surface are rotated versions of each other: the c-axis on the left corresponds to the b-axis on the right.

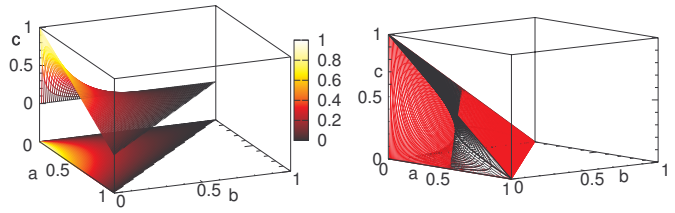


Fig. 7. These pictures characterize the joint distributions $\mathbf{P}^{(X,Y)}$ that allow an ANM in both directions. This is fulfilled if both variables are independent ($ad = bc$, left) or (right) if $\mathbf{P}^{(X,Y)}$ lies on the intersection of the ANM$_{X\to Y}$-surface (black) and the ANM$_{Y\to X}$-surface (red) from Figure 6: $b = c = 0$ corresponds to the a-axis, $a = d = 0$ and thus $c = 1 - b$ to the straight line between $(0, 0, 1)$ and $(0, 1, 0)$ and $a = d, b = c$ (ergo $c = 0.5 - a$) is represented by the intersection line between $(0.5, 0, 0)$ and $(0, 0.5, 0.5)$.

D. Mixed Models

With the results developed in the last two sections we can cover even models with mixed constraints if both variables have finite support. For the precise conditions of "usually" see Theorem 4 in section III-B.

$$
\begin{aligned}
& Y = f(X) + N,\ N \perp\!\!\!\perp X;\ X \text{ cyclic}, Y, N \text{ non-cyclic} \\
\overset{II\text{–}C}{\Rightarrow}\ & Y = f(X) + N,\ N \perp\!\!\!\perp X;\ X \text{ cyclic}, Y, N\ \tilde{m}\text{-cyclic} \\
\overset{\text{Thm } 4}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& X, \tilde{N} \text{ cyclic}, Y\ \tilde{m}\text{-cyclic} \\
\overset{II\text{–}C}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& X, \tilde{N} \text{ cyclic}, Y \text{ non-cyclic}
\end{aligned}
$$

And, conversely:

$$
\begin{aligned}
& Y = f(X) + N,\ N \perp\!\!\!\perp X;\ Y, N \text{ cyclic}, X \text{ non-cyclic} \\
\overset{II\text{–}C}{\Rightarrow}\ & Y = f(X) + N,\ N \perp\!\!\!\perp X;\ Y, N \text{ cyclic}, X\ m\text{-cyclic} \\
\overset{\text{Thm } 4}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& Y \text{ cyclic}, X, \tilde{N}\ m\text{-cyclic} \\
\overset{II\text{–}C}{\Rightarrow}\ & \text{Usually there is no ANM } X = g(Y) + \tilde{N},\ \tilde{N} \perp\!\!\!\perp Y, \\
& Y \text{ cyclic}, X, \tilde{N} \text{ non-cyclic}
\end{aligned}
$$

IV. Practical Method for Causal Inference

Based on our theoretical findings in section III we propose the following method for causal inference (see Hoyer et al. (2009) for the continuous case):

1) Given: iid data of the joint distribution $\mathbf{P}^{(X,Y)}$.
2) Regression of $Y = f(X) + N$ leads to residuals $\hat{N}$, regression of $X = g(Y) + \tilde{N}$ leads to residuals $\hat{\tilde{N}}$.
3) If $\hat{N} \perp\!\!\!\perp X$ and $\hat{\tilde{N}} \not\perp\!\!\!\perp Y$, infer *"$X$ is causing Y"*, if $\hat{N} \not\perp\!\!\!\perp X$ and $\hat{\tilde{N}} \perp\!\!\!\perp Y$, infer *"$Y$ is causing X"*, if $\hat{N} \not\perp\!\!\!\perp X$ and $\hat{\tilde{N}} \not\perp\!\!\!\perp Y$, infer *"I don't know (bad model)"*, if $\hat{N} \perp\!\!\!\perp X$ and $\hat{\tilde{N}} \perp\!\!\!\perp Y$, infer *"I don't know (both directions possible)"*.

(The identifiability results show that the last case will almost never occur.) This procedure requires discrete methods for regression and independence testing and we now discuss our choices. Code is available on the first author's homepage.

A. Regression Method

Given a finite number of iid samples of the joint distribution $\mathbf{P}^{(X,Y)}$ we denote the sample distribution by $\hat{\mathbf{P}}^{(X,Y)}$. In continuous regression we usually minimize a sum consisting of a loss function (like an ℓ_2-error) and a regularization term that prevents us from overfitting.

Regularization of the regression function is not necessary in the discrete case for large sampling. Since we may observe many different values of Y for one specific X value there is no risk in overfitting. This introduces further difficulties compared to continuous regression since in principle we now should try all possible functions from X to Y and compare the corresponding values of the loss function.

Minimizing a *loss function* like an ℓ_p error is not fully appropriate for our purpose, either: after regression we evaluate the proposed function by checking the independence of the residuals. Thus we should choose the function that makes the residuals as independent as possible (see also Mooij et al., 2009). Therefore we consider a dependence measure (DM) between residuals and regressor as loss function, which we denote by $\mathrm{DM}(\hat{N}, X)$.

Two problems remain:
(1) Assume the different X values $x_1 < \ldots < x_n$ occur in the sample distribution $\hat{\mathbf{P}}^{(X,Y)}$. Then one only has to evaluate the regression function on these values. More problematic is the range of the function. Since we can only deal with finite numbers, we have to restrict the range to a finite set. No matter how large we choose this set, it is always possible that the resulting function class does not contain the true function. But since we used the freedom of choosing an additive constant to require $n(0) > n(k)$ and $\tilde{n}(0) > \tilde{n}(k)$ for all $k \neq 0$, we will always find a sample (X_i, Y_i) with $Y_i = f(X_i)$ if the sample size is large enough. Thus it would be reasonable to consider all Y values that occur together with $X = x$ as a potential value for $f(x)$. To even further reduce the impact of this problem we regard *all* values between $\min Y$ and $\max Y$ as possible values for f. And if there are too few samples with $X = x_j$ and the true value $f(x_j)$ is not included in $\{\min Y, \min Y + 1, \ldots, \max Y\}$ we may not find the true function f, but the few "wrong" residuals do not have an impact on the independence. In practice the following second deliberation is more relevant than the first one:
(2) Even if all values of the true function f are one of the $m := \#\{\min Y, \min Y + 1, \ldots, \max Y\}$ considered values, the problem of checking all possible functions is not tractable: If $n = 20$ and $m = 16$ there are $16^{20} = 2^{80}$ possible functions. We thus propose the following heuristic but efficient procedure:

Start with an initial function $f^{(0)}$ that maps every value x to the y which occurred (together with this x) most often under all y. Iteratively we then update each function value separately. Keeping all other function values $f(\tilde{x})$ with $\tilde{x} \neq x$ fixed we choose $f(x)$ to be the value that results in the "most independent" residuals. This is done for all x and repeated up to J times as shown in Algorithm 1. Recall that we required $n(0) \geq n(k)$ for all k.

Algorithm 1 Discrete Regression with Dependence Minimization

Input: $\hat{\mathbf{P}}(X, Y)$
Output: f
$f^{(0)}(x_i) := \operatorname{argmax}_y \hat{\mathbf{P}}(X = x_i, Y = y)$
repeat
  $j = j + 1$
  for i in a random ordering **do**
    $f^{(j)}(x_i) := \operatorname{argmin}_y \mathrm{DM}\big(X, Y - f^{(j-1)}_{x_i \mapsto y}(X)\big)$
  end for
until residuals $Y - f^{(j)}(X) =: \hat{N} \perp\!\!\!\perp X$ **or** $f^{(j)}$ does not change anymore **or** $j = J$.

In the algorithm, $f^{(j-1)}_{x_i \mapsto y}(X)$ means that we use the current version of $f^{(j-1)}$ but change the function value $f(x_i)$ to be y. If the argmax in the initialization step is not unique we take the largest possible y. We can even accelerate the iteration step if we do not consider all possible values $\{\min Y, \ldots, \max Y\}$, but only the five that give the highest values of $\hat{\mathbf{P}}(X = x_i, Y = y)$ instead.

Note that the regression method performs coordinate descent in a discrete space and $\mathrm{DM}\big(X, Y - f^{(j)}(X)\big)$ is monotonically decreasing (and bounded from below). Since $f^{(j)}$ is changed only if the dependence measure can be strictly decreased and furthermore the search space is finite, the algorithm converges towards a local optimum. Although it is not obvious why $f^{(j)}$ should converge towards the *global* minimum, the experimental results will show that the method works very reliably in practice.

B. Independence Test and Dependence Measure

Assume we are given joint iid samples (W_i, Z_i) of the discrete variables W and Z and we want to test whether W and Z are independent. In our implementation we use Pearson's χ^2 test (e.g. Agresti (2002)), which is most commonly used. It computes the difference between observed frequencies and expected frequencies in the contingency table. The test statistic is known to converge towards a χ^2 distribution, which is taken as an approximation even in the finite sample case. In the case of very few samples Cochran (1954) suggests to use this approximation only if more than 80% of the expected counts are larger than 5 ("Cochran's condition"). Otherwise, Fisher's exact test (e.g. Agresti (2002)) could be used. In the remainder of the article we denote the significance level of the test by α.

For a dependence measure DM we use the p-value (times -1) of the independence test. If the p-value is smaller than 10^{-16}, however, it is regarded as 0 and we take the test statistic instead.

V. Experiments

Simulated Data

We first investigate the performance of our method on synthetic data sets. Therefore we simulate data from ANMs and check whether the method is able to rediscover the true model. We showed in section III that only very few examples allow a reversible ANM. Data sets A1 and B1 support these theoretical results. We simulate data from many randomly chosen models. All models that allow an ANM in both directions are instances of our examples from above (without exception). Data sets A2 and B2 show how well our method performs for small data size and models that are close to non-identifiability. Data set A3 empirically investigates the run-time performance of our regression method and compares it with a brute-force search. Data set A4 show that two consecutive ANMs $Z = g(f(X) + N_1) + N_2$ do not necessarily follow a single ANM. Data set B3 shows that the method does not favor one direction if the supports of X and Y are of different size. All experiments are available with the code.

A. Integer Models

Data set A1 (identifiability).
With equal probability we sample from a model with

(1) $\operatorname{supp} X \subset \{1, \ldots, 4\}$
(2) $\operatorname{supp} X \subset \{1, \ldots, 6\}$
(3) X binomial with parameters (n, p)
(4) X geometric with parameter p
(5) X hypergeometric with parameters (M, K, N)
(6) X Poisson with parameter λ or
(7) X negative binomial with parameters (n, p).

For each model the parameters of these distributions are chosen randomly (n, M, K, N uniformly between 1 and $40, 40, M, K$, respectively, p uniformly between 0.1 and 0.9 and λ uniformly between 1 and 10), the functions are random ($f(x) \sim U(\{-7, -6, \ldots, 7\})$ is uniform for each $x \in \operatorname{supp} X$) and the noise distribution is random, too ($S \sim U(\{1, 2, 3, 4, 5\})$ determines the support $\operatorname{supp} N = \{-S, \ldots, S\}$ and $\mathbf{P}^N$ is chosen by drawing $\#\operatorname{supp} N - 1$ numbers in $[0, 1]$ and taking differences). This way we also construct $\mathbf{P}^X$ in cases (1) and (2).

We then consider 1000 different models. For each model we sample 1000 data points and apply our algorithm with a signficance level of $\alpha = 0.05$ for the independence test. The results given in Table I show that the method works well on almost all simulated data sets. The algorithm outputs "bad fit in both directions" in roughly 5% of all cases, which corresponds to the chosen test level. The model is non-identifiable only in 5.3% of the cases, all of which are instances either with a constant function f (2.3%) and thus independent X and Y or with "non-overlapping noise" (3.0%), that is: $f(x) + \operatorname{supp} N$ are disjoint for $x \in X$, which means $\#C_i = 1$ (see Theorem 1). This empirically supports Corollary 2 and therefore our proposition that the model is identifiable in the generic case.

TABLE I
DATA SET A1. THE TRUE DIRECTION IS ALMOST ALWAYS IDENTIFIED.

correct dir.:	89.9%	both dir. poss.:	5.3%
wrong dir.:	0%	bad fit in both dir.:	4.8%

Data set A2 (close to non-identifiable).
For this data set we sample from the model $Y = f(X) + N$ with $n(-2) = 0.2$, $n(0) = 0.5$, $n(2) = 0.3$, and $f(-3) = f(1) = 1$, $f(-1) = f(3) = 2$. Depending on the parameter r we sample X from $p(-3) = 0.1 + r/2$, $p(-1) = 0.3 - r/2$, $p(1) = 0.15 - r/2$,

$p(3) = 0.45 + r/2$. For each value of the parameter r ranging between $-0.2 \leq r \leq 0.2$ we use 100 different data sets, each of which has the size 400. Theorem 1 shows that the ANM is reversible if and only if $r = 0$. Thus, our algorithm does not decide when $r \approx 0$. Figure 8 shows that the algorithm identifies the correct direction for $r \neq 0$. Again, the test level of $\alpha = 5\%$ introduces indecisiveness of roughly the same size, which can be seen for $|r| \geq 0.15$.

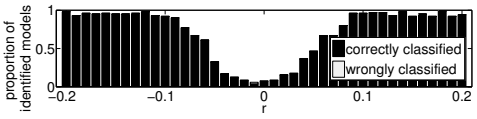


Fig. 8. Data set A2. Proportion of correct and false results of the algorithm depending on the distribution of N. The model is not identifiable for $r = 0$. If r differs significantly from 0 almost all decisions are correct.

Data set A3 (fast regression).
The space of all functions from the domain of X to the domain of Y is growing rapidly in their sizes: If $\#\operatorname{supp} X = m$ and $\#\operatorname{supp} Y = \tilde{m}$ then the space $\mathcal{F} := \{f : \operatorname{supp} X \to \operatorname{supp} Y\}$ has $\tilde{m}^m$ elements. If one of the variables has infinite support the set is even infinitely large (although this does not happen for any finite data set). It is clear that it is infeasible to optimize the regression criterion by trying every single function. As mentioned before one can argue that with high probability it is enough to only check the functions that correspond to an empirical mass that is greater than 0 (again assuming $n(0) > 0$): E.g. it is likely that $\hat{\mathbf{P}}(X = -2, Y = f(-2)) > 0$. We call these functions "empirically supported". But even this approach is often infeasible. In this experiment we compare the number of possible functions (with values between $\min Y$ and $\max Y$), the number of empirically supported functions and the number of functions that were checked by the algorithm we proposed in section IV-A in order to find the true function (which it always did).

We simulate from the model $Y = \operatorname{round}(0.5 \cdot X^2) + N$ for two different noise distributions: $n_1(-2) = n_1(2) = 0.05, n_1(k) = 0.3$ for $|k| \leq 1$ and $n_2(-3) = n_2(3) = 0.05, n_2(k) = 0.18$ for $|k| \leq 2$. Each time we simulate a uniformly distributed X with i values between $-\frac{i-1}{2}$ and $\frac{i-1}{2}$ for $i = 3, 5, \ldots, 19$. For each noise-regressor distribution we simulated 100 data sets. For N_1 and $i = 9$, for example, there are $(11 - (-2))^9 \approx 1.1 \cdot 10^{10}$ possible functions in total and $5^9 \approx 2.0 \cdot 10^6$ functions with positive empirical support. Our method only checked 107 ± 25 functions before termination. The highest number of functions checked by the algorithm is 645 ± 220. The full results are shown in Figure 9.

Data set A4 (limitation of ANMs).
One can imagine that (for a non-linear g) two consecutive ANMs $Z = g(f(X) + N_1) + N_2$ (which could come from a causal chain $X \to Y \to Z$ with unobserved Y) do not necessarily allow an ANM from X to Z. This means that if a relevant intermediate variable is missing, our method would output "I do not know (bad model fit)" and therefore does not propose a causal direction. We hope, however, that even in this situation the joint distribution is often reasonably "closer" to ANM in the correct direction than to an ANM in the wrong direction. We demonstrate this effect on simulated data: We use 300 samples, $\operatorname{supp} X \subset \{1, \ldots, 8\}$ and $\operatorname{supp} N \subset \{-3, \ldots, 3\}$ (the distributions are chosen as in Data set A1), simulated 100 data sets and obtained the results in Table II. Clearly, the effect vanishes if one either increase the sample size (to 2000, say) or one includes even more ANMs between X and Z (results not shown).

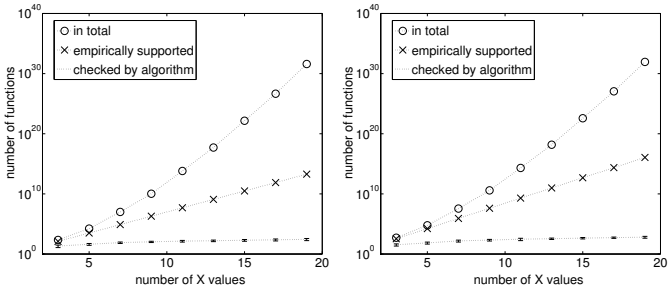


Fig. 9. Data set A3. The size of the whole function space, the number of all functions with empirical support and the number of functions checked by our algorithm (including standard deviation) is shown for N_1 (left) and N_2 (right). An extensive search would be intractable in these cases. The proposed algorithm is very efficient and still finds the correct function for all data sets.

TABLE II
DATA SET A4. SINCE THE DISTRIBUTION DOES NOT ALLOW AN ANM, THE METHOD DOES NOT DECIDE IN MOST CASES. STILL, THE METHOD SEEMS TO PREFER AN ANM IN THE CORRECT DIRECTION.

p-value	$5 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$
correct dir.:	18%	24%	34%	35%
wrong dir.:	5%	4%	2%	5%
both dir. poss.:	2%	18%	27%	36%
bad fit in both dir.:	75%	54%	37%	24%

B. Cyclic Models

Data set B1 (identifiability).
For the three combinations $(m, \tilde{m}) \in \{(3,3), (3,5), (5,3)\}$ we consider 1000 different models each: As in Data Set A1 we randomly choose a function $f \neq \text{const}$, $\mathbf{P}^X$ and $\mathbf{P}^N$. For each model we sample 2000 data points and apply our algorithm with a significance threshold of $\alpha = 0.05$. The results given in Table III show that the method works well on almost all simulated data sets. The algorithm outputs "bad fit in both directions" in roughly 5% of all cases, which corresponds to the chosen test level. The model is non-identifiable only in very few cases. All of these cases are instances of the counter examples 1(i), 1(ii) and 2 from above. Due to space limitations we only show 6 (out of 11) in Table IV. This experiment further supports our theoretical result that the model is identifiable in the generic case.

TABLE III
DATA SET B1. THE ALGORITHM IDENTIFIES THE TRUE CAUSAL DIRECTION IN ALMOST ALL CASES. ONLY IN FEW CASES ANMS CAN BE FIT IN BOTH DIRECTIONS, WHICH SUPPORTS THE RESULTS OF SECTION III.

$(m, \tilde{m})$	$(3,3)$	$(3,5)$	$(5,3)$
correct dir.:	95.3%	94.8%	95.5%
wrong dir.:	0.0%	0.0%	0.0%
both dir. poss.:	0.8%	0.0%	0.3%
bad fit in both dir.:	3.9%	5.2%	4.2%

TABLE IV

DATA SET B1. THIS TABLE SHOWS ONLY SOME CASES, WHERE ANMS IN BOTH DIRECTIONS WERE POSSIBLE. ALL CASES (INCLUDING THE ONES NOT SHOWN) ARE INSTANCES OF THE EXAMPLES GIVEN IN SECTION III.

Function f	$p(1), \dots, p(m)$	$n(1), \dots, n(\tilde{m})$	Instance of Example
$0 \mapsto 0, 1 \mapsto 2, 2 \mapsto 0$	0.83, 0.00, 0.17	$0.15, 0.26, 0.58$	$1(i)$
$0 \mapsto 2, 1 \mapsto 0, 2 \mapsto 2$	0.34, 0.53, 0.14	$0.33, 0.34, 0.33$	$1(ii)$
$0 \mapsto 2, 1 \mapsto 1, 2 \mapsto 0$	0.33, 0.33, 0.34	$0.85, 0.14, 0.02$	2
$0 \mapsto 1, 1 \mapsto 0, 2 \mapsto 1, 3 \mapsto 0, 4 \mapsto 0$	0.20, 0.47, 0.14, 0.08, 0.12	$0.33, 0.33, 0.34$	$1(ii)$
$0 \mapsto 1, 1 \mapsto 0, 2 \mapsto 1, 3 \mapsto 1, 4 \mapsto 1$	0.55, 0.01, 0.03, 0.26, 0.14	$0.37, 0.32, 0.31$	$1(i)$
$0 \mapsto 0, 1 \mapsto 1, 2 \mapsto 0, 3 \mapsto 1, 4 \mapsto 2$	0.03, 0.71, 0.06, 0.10, 0.32	$0.32, 0.34, 0.34$	$1(ii)$

Data set B2 (close to non-identifiable).
For this data set let $m = \tilde{m} = 4$ and $f = \mathrm{id}$. The distribution of p is given by: $p(0) = 0.6, p(1) = 0.1, p(2) = 0.1, p(3) = 0.2$. Depending on the parameter $\frac{1}{2} \leq r \leq \frac{4}{5}$ we sample the noise N from the distribution $n(0) = n(1) = r/2, n(2) = n(3) = 1/2 - r/2$. That means N is uniformly distributed if and only if $r = \frac{1}{2}$. Thus, the model is not identifiable if and only if the noise distribution is uniform, i.e. if and only if $r = \frac{1}{2}$.
(This can be seen as follows: Since $\mathbf{P}(X = 0, Y = 0) > \mathbf{P}(X = k, Y = 0)$ and $\mathbf{P}(X = 0, Y = 1) > \mathbf{P}(X = k, Y = 1)$ for all $k \neq 0$ we have that $g(0) = 0 = g(1)$, still assuming $\mathbf{P}(\tilde{N} = 0) > \mathbf{P}(\tilde{N} = k)$ for all $k \neq 0$. Thus g is not injective. The special form of f leads to one cycle of length 4, which implies that uniformly distributed N is a necessary condition for a reversible ANM, see Proposition 7 in section VI. Example $1(ii)$ shows that it is also sufficient.)
The further r is away from $\frac{1}{2}$, the easier it should be for our method to detect the true direction. For each value of the parameter r we use 100 different samples, each of which has size 200. This time we choose a significance level of 0.01, which still leads to no wrong decisions (see Figure 10).

For $r = 0.58$ and $r = 0.68$ (indicated by the arrows in Figure 10) we further investigate the dependence on the data size. Clearly, $r = 0.58$ results in a model that is still very close to non-identifiability and thus we need more data to perform well, whereas for $r = 0.68$ the performance increase quickly with the sample size (see Figure 11). Note that non-identifiable models lead to very few, but not to wrong decisions.

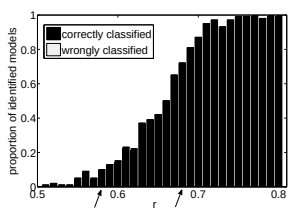


Fig. 10. Data set B2. Proportion of correct results of the algorithm depending on the distribution of N. The model is not identifiable for $r = 0.5$. If r differs significantly from 0.5 the algorithm makes a decisions in almost all cases.

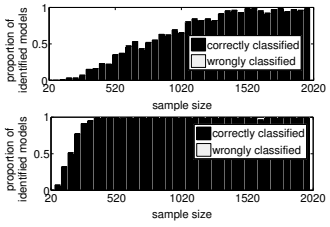


Fig. 11. Data Set B2. For $r = 0.58$ (top) and $r = 0.68$ (bottom) the performance depending on the data size is shown. More data is needed if the true model is close to non-identifiable (top). In both cases the performance clearly increases with the sample size.

Data set B3 (no direction is favored a priori).
Here, we consider two random variables, which supports are very unequal in size. If we choose $m := \#\mathcal{X} := \#\mathrm{supp}\, X = 2$ and $\tilde{m} := \#\mathcal{Y} := \#\mathrm{supp}\, Y = 10$, there are $2^{10} = 1024$ function from $\mathcal{Y}$ to $\mathcal{X}$, but only $10^2 = 100$ functions from $\mathcal{X}$ to $\mathcal{Y}$; one could expect the method to favor models from Y to X. We show that this is not the case.

For $m \neq \tilde{m} \in \{2, 10\}$ and $m \neq \tilde{m} \in \{3, 20\}$ we randomly choose distributions for X and N and a function f (as before) and sampled 500 data points from this forward ANM. Table V shows that the algorithm detects the true direction in almost all cases (except if the model is non-identifiable).

TABLE V

DATA SET B3. THE ALGORITHM IDENTIFIES THE TRUE CAUSAL DIRECTION IN ALMOST ALL CASES. THERE IS NO EVIDENCE THAT THE ALGORITHM ALWAYS FAVORS ONE DIRECTION.

m	$\tilde{m}$	cor. dir.	wrong dir.	both dir. poss.	both dir. bad fit
2	10	97.4%	0%	2.5%	0.1%
10	2	85.2%	0%	14.8%	0.0%
3	20	96.8%	0%	1.6%	1.6%
20	3	95.5%	0%	4.4%	0.1%

Real Data.

Data set 5 (abalone).
We also applied our method to the `abalone` data set (Nash et al., 1994) from the UCI Machine Learning Repository (Asuncion & Newman, 2007). We tested the sex X of the abalone (male (1), female (2) or infant (0)) against length Y_1, diameter Y_2 and height Y_3, which are all measured in mm, and have $70, 57$ and 28 different values, respectively. Compared to the number of samples (up to 4177) we treat this data as being discrete. Because we do not have information about the underlying continuous length we have to assume that the data structure has not been destroyed by the user-specific discretization. We regard $X \to Y_1$, $X \to Y_2$ and $X \to Y_3$ as being the ground truth, since the sex is probably causing the size of the abalone, but not vice versa.

Clearly, the Y variables do not have a cyclic structure. For the sex variable, however, the most natural model would be a structureless set which is contained in the cyclic constraints; for comparison we try both models for X. Our method is able to identify 2 out of 3 directions correctly and does not make a decision in one case. Except for this one exception all of the backward models are rejected (see Table VI and Figure 12). We used the test level $\alpha = 5\%$ and the first 1000 samples of the data set.

For this data set the method proposed by (Sun et al., 2006) returns a slightly higher likelihood for the true causal directions

TABLE VI

DATA SET 5. THE ALGORITHM IDENTIFIES THE TRUE CAUSAL DIRECTION IN 2 CASES. ALSO FOR Y_1 THE p-VALUE IS HIGHER FOR THE CORRECT DIRECTION, BUT FORMALLY THE METHOD DOES NOT MAKE A DECISION. HERE, WE ASSUMED A NON-CYCLIC STRUCTURE ON Y AND TRIED BOTH CYCLIC AND NON-CYCLIC FOR X.

	Y_1	Y_2	Y_3
p-value$_{X\to Y}$	0.17	0.19	0.05
p-value$_{Y\to X}$ (non-cyclic)	$6\cdot 10^{-12}$	$2\cdot 10^{-14}$	$<10^{-16}$
p-value$_{Y\to X}$ (cyclic)	0.06	$4\cdot 10^{-3}$	$1\cdot 10^{-8}$

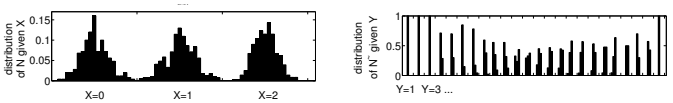


Fig. 12. Data set 5. For Y_3 regressing on X (left) and vice versa (right) the plot shows the conditional distribution of the fitted noise given the regressor. If the noise is independent, then the distribution must not depend on the regressor state. Clearly, this is only the case for $X\to Y_3$ (left), which corresponds to the ground truth.

than for the false directions, but this difference is so small, that their algorithm does not consider it to be significant.

The abalone data set also shows that working with p-values requires some care. For synthetic data sets that we simulate from one fixed model the p-values do not depend on the data size. In real world data, however, this often is the case. If the data generating process does not exactly follow the model we assume, but is reasonable close to it, we get good fits for moderate data sizes. Only including more and more data reveals the small difference between process and model, which therefore leads to small p-values. Figure 13 shows how the p-values vary if we include the first n data points of the abalone data set (in total: 4177). One can see that although the p-values for the correct direction decrease they are clearly preferable to the p-values of the wrong direction. This is a well-known problem in applied statistics that also has to be considered using our method.

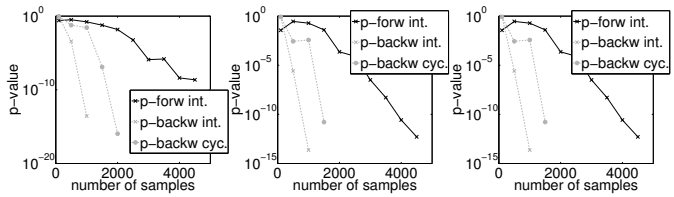


Fig. 13. Data set 5. The plots show p-values of forward and backward direction depending on the number of samples we included (no data point means $p<10^{-16}$). The p-value in the correct direction is eventually lower than any reasonable threshold. Nevertheless we prefer this direction since it is decreasing much more slowly than p-backward.

Data set 6 (acute inflammations).
The data set `acute inflammations` (Czerniak & Zarzycki, 2003) from the UCI ML Repository (Asuncion & Newman, 2007) consists of 120 patients. For each patient we have an indicator that tells us whether a specific symptom is present or absent, the temperature and the diagnosis of a medical expert, whether the patient suffers from acute inflammations of urinary bladder and/or whether he suffers from acute nephritises. In particular, we have binary indicators (yes or no) Y_1: occurrence of nausea, Y_2: lumbar pain, Y_3: urine pushing, Y_4: micturition pains and Y_5: burning of urethra, itch, swelling of urethra outlet. Furthermore, the temperature T is measured in $^\circ C$ with $0.1^\circ C$ accuracy. We denote the diagnosis by X_1 (inflammation of urinary bladder) and X_2 (nephritis of renal pelvis origin).

Since the medical expert's diagnosis is based only on the symptoms we expect $Y\to X_i$ and $T\to X_i$ for $i=1,2$ (precisely, we expect all Y's and T to be *common* causes for X_i, but here, we only consider the bivariate case and hope that the method still works). It is crucial that the variables X_i only indicate the *diagnosis* and not necessarily the truth. If the X_i corresponded to the true state, X_i would be regarded as the cause and Y as the effect. But in this data set we model the diagnosis behavior of doctors and not the disease process in the patients.

Note further that except for T all variables are binary and should be modeled as being cyclic. The results are presented in Table VII. Since T takes 44 different values and the sample size is only 120 we also introduce $T_*:=\text{round}(T)$ that only takes 7 values. This is necessary in order to meet Cochran's condition and get reliable results from the independence test. (We are aware that on the other hand, this may introduce small changes in the data generating model, but we hope that this has no effect on the causal reasoning.) The method correctly identifies the causal

TABLE VII

DATA SET 6. THE ALGORITHM IDENTIFIES THE TRUE CAUSAL DIRECTION IN FOUR CASES (BOLD FONT). IN ALL OTHER CASES THE METHOD DOES NOT DECIDE. THE ASTERISKS INDICATE, WHERE ONE p-VALUE IS AT LEAST 10 TIMES LARGER THAN THE OTHER.

	p-val$_{X_1\to Y}$	p-val$_{Y\to X_1}$	p-val$_{X_2\to Y}$	p-val$_{Y\to X_2}$
Y_1	0.043	**0.368**	$2\cdot 10^{-9}$	$0.004\,*$
Y_2	0.043	**0.368**	$3\cdot 10^{-5}$	$3\cdot 10^{-5}$
Y_3	$7\cdot 10^{-7}$	$4\cdot 10^{-4}\,*$	0.009	0.009
Y_4	0.935	0.935	0.925	0.102
Y_5	0.102	0.925	0.190	0.190
T	0.556	1.000	0.080	$0.997\,*$
T_*	0.013	**0.435**	0.005	**0.142**

links $Y_1\to X_1$, $Y_2\to X_1$, $T_*\to X_1$ and $T_*\to X_2$. In six more cases the method does not decide. This is relatively often and may be explained by the small data size, for which it is difficult to reject a null hypothesis. We therefore assign an asterisks to all further directions for which the corresponding p-value is at least 10 times larger than the one for the other direction. Furthermore, we checked that the method does not find any causal link between the symptom variables Y, as expected.

Here, the method from (Sun et al., 2006) does not find a significant result in 12 cases (8 cases: exactly the same likelihood for both directions, 3 cases: small favor of the wrong direction, 1 case: small favor for the correct direction) and it wrongly infers $X_2\to T$ and $X_2\to T_*$ as being significant.

Data set 7 (temperature).
We further applied our method to a data set consisting of 9162 daily values of temperature measured in Furtwangen (Germany)[1] using the variables temperature (T, in $^\circ C$) and month (M). Clearly M inherits a cyclic structure, whereas T does not. Since the month indicates the position of the earth relatively to the sun,

[1]B. Janzing contributed this data set. It is one pair on `https://webdav.tuebingen.mpg.de/cause-effect/`

which is causing the temperature on earth, we take $M \to T$ as the ground truth. Here, we aggregate states and use months instead of days. Again, this is done in order to meet Cochran's condition; it is not a scaling problem of our method (if we do not aggregate the method returns $p_{\text{days}\to T} = 0.9327$ and $p_{T\to\text{days}} = 1.0000$).

For 1000 data points both directions are rejected (p-value$_{M\to T} = 3e-4$, p-value$_{T\to M} = 1e-13$). Figure 14 shows, however, that again the p-values$_{M\to T}$ are decreasing much more slowly than p-values$_{T\to M}$. Using other criteria than simple p-values we still may prefer this direction and propose it as the true one.

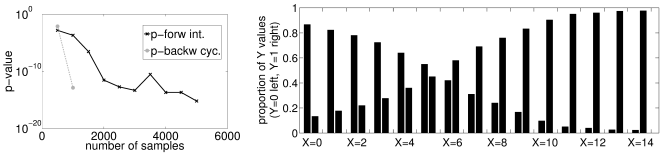


Fig. 14. Data set 7 and Data set 8. Left: The plot show p-values of forward and backward direction depending on the number of samples we included (no data point means $p < 10^{-16}$). Again we prefer the correct direction since the p-values are decreasing much more slowly than p-backward. Right: The data set does not allow an ANM in any of the two directions. Therefore the method does not propose an answer.

The method proposed in (Sun et al., 2006) finds a larger likelihood for the correct direction, but does not consider this difference as being significant.

Data set 8 (faces).
This data set (Armann & Bülthoff, 2010) (4499 instances) shows the limitations of ANMs. Here, X represents a parameter used to create pictures of artificial faces. X takes values between 0 and 14, where 0 corresponds to a female face, 14 corresponds to a face that is rather masculine. All other parameter values are interpolated. These faces were shown to some subjects who had to indicate whether they believe this is a male ($Y = 1$) or a female ($Y = 0$) face. In this example we regard X as being the cause of Y. However, the data do not admit an ANM in any direction ($p_{X\to Y} = 0$ and $p_{Y\to X} = 0$). Thus, the method does not make a mistake, but does not find the correct answer, either. On this data set the method in (Sun et al., 2006) again detects an insignificantly larger likelihood for the correct direction.

It is possible, however, to include generalizations of ANMs that are capable of modeling this data set. One possibility is to consider models of the form

$$Y = f(X+N), N \perp\!\!\!\perp X \quad \text{and} \quad X = g(Y+\tilde{N}), \tilde{N} \perp\!\!\!\perp Y \quad (3)$$

with some possibly non-invertible functions f and g (for continuous data, a similar model has been proposed by Zhang & Hyvarinen (2009)). In this model the function f does not only act on the support of X, but on an enlarged space. Using a method that is based on the same ideas described in section IV one is able to fit this data set quite well ($p_{X\to Y} = 1.000$ and $p_{Y\to X} = 0$). However, we do not have any theoretical identifiability results and the method has one further drawback: Simulations show that it often prefers the variable with the smaller support as the effect.

In particular, we can indicate why the model class at the right hand side of equation (3) gets too large if X is a binary variable and Y is the discretization of a continuous variable: If one sets g to be the Heaviside step function defined by $g(w) = 1$ if $w \geq 0$ and $g(w) = 0$ otherwise, equation (3) leads to (with $m(t)$ the probability mass function of $M := -\tilde{N}$)

$$\mathbf{P}(X = 1|Y = y) = \mathbf{P}(y + \tilde{N} \geq 0) = \mathbf{P}(M \leq y) = \sum_{t=-\infty}^{y} m(t).$$

Hence, every conditional for which $\mathbf{P}(X = 1|Y = y)$ is monotonously increasing can be described by an ANM. But even some models that we regard as a natural examples for $X \to Y$ lead to such a monotonously increasing conditional: E.g. $\mathbf{P}^{Y|X=0}$ and $\mathbf{P}^{Y|X=1}$ being discretized Gaussians with equal variance and different means.

VI. Proofs

A. Proof of Theorem 1

Proof:

$\Rightarrow$: First we assume $\operatorname{supp} Y = \{y_0, \ldots, y_m\}$ with $y_0 < y_1 < \ldots < y_m$. This implies that $N_{\max} := \min\{n \in \mathbb{N} \,|\, \mathbf{P}(N = n) > 0\}$ is finite. Define the non-empty sets $\tilde{C}_i := \operatorname{supp} X|Y = y_i$, for $i = 0, \ldots, m$. That means $\tilde{C}_0, \ldots, \tilde{C}_m \subset \operatorname{supp} X$ are the smallest sets satisfying $\mathbf{P}(X \in \tilde{C}_i \,|\, Y = y_i) = 1$. For all i, j it follows that

$$\tilde{C}_i = \tilde{C}_j \text{ or } \tilde{C}_i \cap \tilde{C}_j = \emptyset \text{ and } f\,|_{\tilde{C}_i} = \tilde{c}_i = \text{const.} \quad (4)$$

This is proved by an induction argument.

Base step: Consider $\tilde{C}_m$ corresponding to the largest value $y_m = \max\{f(x) \,|\, x \in X\} + N_{\max}$ of $\operatorname{supp} Y$. Assuming $f(x_1) < f(x_2)$ for $x_1, x_2 \in \tilde{C}_m$ leads to $y_m = f(x_1) + N_{\max} < f(x_2) + N_{\max} = y_m$ and therefore to a contradiction. Induction step: Consider $\tilde{C}_k$ and assume properties (4) are satisfied for all $\tilde{C}_{\tilde{k}}$ with $k < \tilde{k} \leq m$. If $x \in \tilde{C}_k \cap \tilde{C}_{\tilde{k}}$ for some $\tilde{k}$

$$\Rightarrow \mathbf{P}(N = y_k - f(\tilde{x})) = \mathbf{P}(N = y_k - f(x)) > 0 \quad \forall \tilde{x} \in \tilde{C}_{\tilde{k}}$$
$$\Rightarrow \tilde{C}_{\tilde{k}} \subset \tilde{C}_k \quad \Rightarrow \tilde{C}_{\tilde{k}} = \tilde{C}_k \quad \Rightarrow f\,|_{\tilde{C}_k} = f\,|_{\tilde{C}_{\tilde{k}}} = \text{const}$$

Furthermore, if $\tilde{C}_k \cap \tilde{C}_{\tilde{k}} = \emptyset \forall k < \tilde{k} \leq m$, then $f\,|_{\tilde{C}_k} = \text{const}$ using the same argument as for C_m.

Thus we can choose some sets $C_0, \ldots, C_l$ from $\tilde{C}_0, \ldots, \tilde{C}_m$, where $l \leq m$, such that $C_0, \ldots, C_l$ are disjoint, and $c_k := f(C_k)$ are pairwise different values. Without loss of generality assume $C_0 = \tilde{C}_0$. Further, even the sets

$$c_k + \operatorname{supp} N := \{c_k + h : \mathbf{P}(N = h) > 0\}$$

are pairwise different: If $y_i = c_k + h_1 = c_l + h_2$ then $C_k \subset \operatorname{supp}(X|Y = y_i) = \tilde{C}_i$ and $C_l \subset \tilde{C}_i$, which implies $k = l$. Now consider the case where Y has infinite and X finite support: $\operatorname{supp} X = \{x_0, \ldots, x_p\}$. Then we define $C_0, \ldots, C_l$ to be disjoint sets, such that f is constant on each of them: $c_i := f(C_i)$. This time, it does not matter which of these sets is called C_0. Again, we will deduce that the sets $c_k + \operatorname{supp} N$ are disjoint:

The sets $\tilde{D}_i := \operatorname{supp} Y|X = x_i$ fulfill

$$\tilde{D}_i = \tilde{D}_j \text{ or } \tilde{D}_i \cap \tilde{D}_j = \emptyset \text{ and } g\,|_{\tilde{D}_i} = \tilde{d}_i = \text{const.}$$

Thus we have $c_k + \operatorname{supp} N$ and $c_k + \operatorname{supp} N$ are either equal or disjoint. But if $c_k + \operatorname{supp} N = c_l + \operatorname{supp} N$ for

$k \neq l$ it follows for $x_a \in C_k$, $x_b \in C_l$ and all $y \in c_k + \operatorname{supp} N$ (since there is a backward model $X = g(Y)+\tilde{N}$)

$$\begin{aligned}&\frac{\mathbf{P}(X = x_a, Y = y)}{\mathbf{P}(X = x_b, Y = y)} = const\\ \Rightarrow\ &\frac{\mathbf{P}(X = x_a) \cdot \mathbf{P}(N = y - f(x_a))}{\mathbf{P}(X = x_b) \cdot \mathbf{P}(N = y - f(x_b))} = const\\ \Rightarrow\ &\frac{\mathbf{P}(N = y - f(x_a))}{\mathbf{P}(N = y - f(x_b))} = const\end{aligned}$$

and thus $\mathbf{P}(N = 0)/\mathbf{P}(N = r) = const, \forall r \in \operatorname{supp} N$. This is only possible for a uniformly distributed N, which leads to a contradiction since Y has been assumed to have infinite support.

Thus we have proved condition c). For a) it remains to show that the sets C_i are shifted versions of each other. This part of the proof is valid for both cases (either X or Y has finite support): Consider C_i for any i. According to the assumption that an ANM $Y \to X$ holds we have

$$\begin{aligned}&\tilde{N}|Y = c_0 \stackrel{d}{=} \tilde{N}|Y = c_i\\ \Leftrightarrow\quad &X - g(c_0)|Y = c_0 \stackrel{d}{=} X - g(c_i)|Y = c_i\\ \Rightarrow\quad &X + d_i|Y = c_0 \stackrel{d}{=} X|Y = c_i \qquad (*)\end{aligned}$$

with $d_i = g(c_i) - g(c_0)$. Thus $C_i = C_0 + d_i$ (including $d_0 = 0$), which completes conditions a).

To prove b) observe that we have for all $x \in C_i$

$$\begin{aligned}\frac{\mathbf{P}(X = x)}{\mathbf{P}(X \in C_i)} &= \frac{\mathbf{P}(X = x)\mathbf{P}(N = c_i - f(x))}{\sum_{\tilde{x}\in C_i} \mathbf{P}(X = \tilde{x})\mathbf{P}(N = c_i - f(\tilde{x}))}\\ &= \frac{\mathbf{P}(X = x, N = c_i - f(x))}{\mathbf{P}(Y = c_i)} = \mathbf{P}(X = x \,|\, Y = c_i)\\ &\stackrel{(*)}{=} \mathbf{P}(X = x - d_i \,|\, Y = c_0)\\ &= \frac{\mathbf{P}(X = x - d_i, N = c_0 - f(x - d_i))}{\mathbf{P}(Y = c_0)}\\ &= \frac{\mathbf{P}(X = x - d_i)}{\mathbf{P}(X \in C_0)}\end{aligned}$$

$\Leftarrow$: In order to show that we have a reversible ANM, we have to construct a g, such that $X = g(Y) + \tilde{N}$. Therefore define the function g as follows: $g(y) = 0, \forall y \in c_0 + \operatorname{supp} N$ and $g(y) = d_i, \forall y \in c_i + \operatorname{supp} N$, $i > 0$. (This is well-defined because of a) and c).) The noise $\tilde{N}$ is determined by the joint distribution $\mathbf{P}^{(X,Y)}$, of course. It remains to check, whether the distribution of $\tilde{N}|Y = y$ is independent of y. Consider a fixed y and choose i such that $y \in c_i + \operatorname{supp} N$. Since $C_i = C_0 + d_i$ the condition $g(y) + h \in C_i$ is satisfied for all $h \in C_0$ and therefore independently of y and c_i. Now, if $g(y) + h \in C_i$ we have

$$\begin{aligned}\mathbf{P}(\tilde{N} = h \,|\, Y = y) &= \frac{\mathbf{P}(X = g(y) + h, Y = y)}{\mathbf{P}(Y = y)}\\ &= \frac{\mathbf{P}(X = g(y) + h, N = y - f(g(y) + h))}{\mathbf{P}(Y = y)}\\ &= \frac{\mathbf{P}(X = g(y) + h)\mathbf{P}(N = y - c_i)}{\sum_{\tilde{x}\in C_i} \mathbf{P}(X = \tilde{x})\mathbf{P}(N = y - f(\tilde{x}))}\\ &= \frac{\mathbf{P}(X = g(y) + h)}{\mathbf{P}(X \in C_i)} = \frac{\mathbf{P}(X = g(y) + h - d_i)}{\mathbf{P}(X \in C_0)}\\ &= \frac{\mathbf{P}(X = h)}{\mathbf{P}(X \in C_0)}\end{aligned}$$

which does not depend on y. And if $g(y) + h \notin C_i$ then $\mathbf{P}(\tilde{N} = h \,|\, Y = y) = 0$, which does not depend on y either.

■

B. Proof of Theorem 3

Proof: We distinguish between two different cases:

a) $\mathbf{P}(N = k) > 0 \,\forall\, m \leq k \leq l$ and $\mathbf{P}(N = k) = 0$ for all other k.

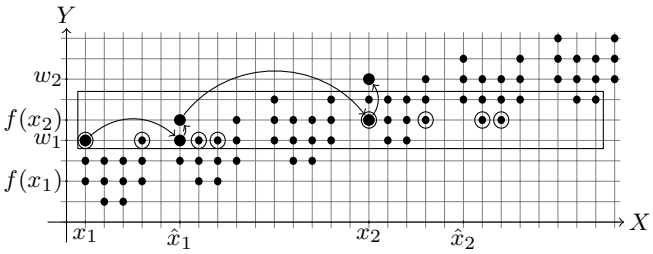


Fig. 15. Visualization of the path from equation (5). Here, $w_2 = f(x_2) + N_{\max}$ and $w_1 = f(x_1) + N_{\max}$.

$\Rightarrow$: Assume that there is an ANM in both directions $X \to Y$ and $Y \to X$. As mentioned above we have a freedom of choosing an additive constant for the regression function. In the remainder of this proof we require $\mathbf{P}(N = k) = \mathbf{P}(\tilde{N} = k) = 0 \,\forall k < 0$ and $\mathbf{P}(\tilde{N} = 0), \mathbf{P}(N = 0) > 0$. The largest k, such that $\mathbf{P}(N = k) > 0$ will be called $N_{\max}$. In analogy to the proof above we define $C_y := \operatorname{supp} X|Y = y$ for all $y \in \operatorname{supp} Y$.

At first we note that all C_y are shifted versions of each other (since there is a backward ANM) and additionally, they are finite sets (otherwise it follows from the compact support of N that there are infinitely many infinite sets $f^{-1}(f(x))$ on which f is constant, which contradicts the assumptions.) Start with any x_1 that satisfies $x_1 = \min\{f^{-1}(f(x_1))\}$ and define

$$\hat{x}_1 := \min\left\{x \in C_{f(x_1)+N_{\max}} \setminus f^{-1}(f(x_1))\right\}$$

This implies $f(\hat{x}_1) > f(x_1)$ and $x_1 \in C_{f(\hat{x}_1)}$.

If such an $\hat{x}_1$ does not exist because the set on the right hand side is empty, then it cannot exist for any choice of x_1: It is clear that $C_{f(x_1)+N_{\max}} = f^{-1}(f(x_1))$ and then we consider the first $C_{f(x_1)+N_{\max}+i}$ that is not empty. Then this set must be $f^{-1}(f(\hat{x}_1))$ for some $\hat{x}_1$. This leads to an iterative procedure and to the required decomposition of $\operatorname{supp} X$.

We have that either

$$\begin{aligned}&\max\{f^{-1}(f(\hat{x}_1))\} > \max\{f^{-1}(f(x_1))\} \text{ or}\\ &\min\{f^{-1}(f(\hat{x}_1))\} < \min\{f^{-1}(f(x_1))\} :\end{aligned}$$

Otherwise $C_{f(\hat{x}_1)}$ and $C_{f(\hat{x}_1)-1}$ satisfy $\max C_{f(\hat{x}_1)-1} \geq \max C_{f(\hat{x}_1)}$ and $\min C_{f(\hat{x}_1)-1} \leq \min C_{f(\hat{x}_1)}$. Because of $\hat{x}_1 \in C_{f(\hat{x}_1)}, \hat{x}_1 \notin C_{f(\hat{x}_1)-1}$ this contradicts the existence of an backward ANM. We therefore assume without loss of generality $\max\{f^{-1}(f(\hat{x}_1))\} > \max\{f^{-1}(f(x_1))\}$. Then we even have $\hat{x}_1 > x_1$, $x_1 = \min\{C_{f(x_1)+N_{\max}}\}$ and $\hat{x}_1 = \min\{C_{f(x_1)+N_{\max}+1}\}$. (Otherwise we use the

same argument as above with $C_{f(x_1)+N_{\max}}$ and $C_{f(x_1)+N_{\max}+1}$.) Define further

$$x_2 := \min f^{-1}(f(x_1) + N_{\max} + 1)$$

Since $f^{-1}(f(x_1)) \subset C_{f(x_1)+N_{\max}}$, but $f^{-1}(f(x_1)) \cap C_{f(x_1)+N_{\max}+1} = \emptyset$, such a value must exist. Again, we can define $\hat{x}_2$ in the same way as above.
Set $y_1 := f(x_1) + N_{\max}$ and $z_1 := f(x_1) + 2 \cdot N_{\max}$ and consider the finite box from $(\min C_{y_1}, y_1)$ to $(\max C_{z_1}, z_1)$. This box contains all the support from $X \,|\, Y = f(x_1) + N_{\max} + i$, where $i = 0, \dots, N_{\max}$. Assume we know the positions in this box, where $\mathbf{P}^{(X,Y)}$ is larger than zero. Then this box determines the support of $X \,|\, Y = f(x_1) + 2 \cdot N_{\max} + 1$ (the line above the box) just using the support of N and $\tilde{N}$. Iterating gives us the whole support of $\mathbf{P}^{(X,Y)}$ in the box above (from $y_2 = f(x_2) + N_{\max}$ to $z_2 = f(x_2) + 2 \cdot N_{\max}$). Since the width of the boxes are bounded by $3 \cdot \max C_{f(x_1)} - \min C_{f(x_1)}$, for example, at some point the box of x_n must have the same support as the one of x_1. Figure 15 shows an example, in which $n = 2$. Using only the distributions of N and $\tilde{N}$ we can now determine a factor α for which $\mathbf{P}(X = x_1, Y = f(x_1) + N_{\max}) = \alpha \cdot \mathbf{P}(X = x_n, Y = f(x_n) + N_{\max})$ This is done by following a sequence between (x_1, y_1) and (x_n, y_n) using only horizontal and vertical steps:

$$(x_1, y_1), (\hat{x}_1, y_1), (\hat{x}_1, f(x_2)), (x_2, f(x_2)), \\ (x_2, y_2), (\hat{x}_2, y_2), \dots, (x_n, y_n) \quad (5)$$

(cf Figure 15). Since this factor only depends on the distributions of N and $\tilde{N}$, the same α satisfies $\mathbf{P}(X = x_n, Y = f(x_n) + N_{\max}) = \alpha \cdot \mathbf{P}(X = x_{2n-1}, Y = f(x_{2n-1}) + N_{\max})$ and therefore

$$\mathbf{P}(X = x_1, Y = f(x_1) + N_{\max}) = \alpha^k \cdot \\ \mathbf{P}(X = x_{(k+1)n-k}, Y = f(x_{(k+1)n-k}) + N_{\max})$$

Note that a corresponding equation with the same constant α holds for the direction to the left of x_1. This leads to a contradiction, since there is no probability distribution for X with infinite support that can fulfill this condition (no matter if α is greater, equal or smaller than 1).

$\Leftarrow$: This direction is proved in exactly the same way as in Theorem 1.

b) $\mathbf{P}(N = k) > 0 \,\forall\, k \in \mathbb{Z}$.
Since X and Y are dependent there are y_1 and y_2, such that $g(y_1) \neq g(y_2)$ with g being the "backward function". Comparing $\{\mathbf{P}(X = k, Y = y_1), k \geq m\}$ and $\{\mathbf{P}(X = k, Y = y_2), k \geq m\}$ we can identify the difference $d := g(y_2) - g(y_1)$. Wlog consider $d > 0$. We use $\frac{\mathbf{P}(X=m-1,Y=y_1)}{\mathbf{P}(X=m,Y=y_1)} = \frac{\mathbf{P}(X=m+d-1,Y=y_2)}{\mathbf{P}(X=m+d,Y=y_2)}$ in order to determine $\mathbf{P}(X = m-1, Y = y_1)$ and then $\mathbf{P}(X = m-1)$ (using f and $\mathbf{P}^N$). Iterations lead to all $\mathbf{P}(X = x)$.

■

C. Proof of Theorem 4

Each distribution $Y \,|\, X = x_j$ has to have the same support (up to an additive shift) and thus the same number of elements with probability larger than 0: $\#\mathrm{supp}\, X \cdot \#\mathrm{supp}\, N = k \cdot \#\mathrm{supp}\, Y$. This proofs (i). For (ii) we now consider 3 different cases: 1. f and g are bijective, 2. g is not injective and 3. f is not injective. These three cases are sufficient since f and g injective implies $n = m$ and f and g bijective. For each of those cases we show that a necessary condition for reversibility includes at least one additional equality constraint for $\mathbf{P}^X$ or $\mathbf{P}^N$.

1st case: f and g are bijective.
Proposition 6: Assume $Y = f(X) + N, N \perp\!\!\!\perp X$ for bijective f and $n(l) \neq 0, p(k) \neq 0 \,\forall k, l$. If the model is reversible with a bijective g, then X and Y are uniformly distributed.

Proof: Since g is bijective we have that $\forall y \exists t_y : g(t_y) = g(y) - 1$. From (2) we can deduce

$$\frac{n\big(y - f(x+1)\big)p(x+1)}{n\big(t_y - f(x)\big)p(x)} = \frac{\tilde{n}\big(x + 1 - g(y)\big)q(y)}{\tilde{n}\big(x + 1 - g(y)\big)q(t_y)}$$

which implies

$$\frac{p(x+1)}{p(x)} = \frac{n\big(t_y - f(x)\big)q(y)}{n\big(y - f(x+1)\big)q(t_y)} \quad \text{and}$$

$$1 = \frac{p(x+m)}{p(x)} = \frac{\prod_{k=0}^{m-1} n\big(t_y - f(x+k)\big)q(y)^m}{\prod_{k=0}^{m-1} n\big(y - f(x+k+1)\big)q(t_y)^m}$$

Since f is bijective it follows that $q(y) = q(t_y)$. This holds for all y and thus Y and X are uniformly distributed. ■

2nd case: g is not injective.
Assume $g(y_0) = g(y_1)$. From (2) it follows that $\frac{n\left(y_0 - f(x)\right)}{n\left(y_1 - f(x)\right)} = \frac{q(y_0)}{q(y_1)} \,\forall x$ and thus $\frac{n\left(y_0 - f(x)\right)}{n\left(y_1 - f(x)\right)} = \frac{n\left(y_0 - f(\tilde{x})\right)}{n\left(y_1 - f(\tilde{x})\right)} \,\forall x, \tilde{x}$, which imply equality constraints on n.
To determine the number of constraints we define a function that maps the arguments of the numerator to those of the denominator

$$h_{y_0,y_1,f} : \begin{array}{rcl} \mathrm{Im}(y_0 - f) & \to & \mathbb{Z}/\tilde{m}\mathbb{Z} \\ y_0 - f(x) & \mapsto & y_1 - f(x) \end{array} .$$

We say h has a cycle if there is a $z \in \mathbb{N}$, s.t. $h^k(a) = (h \circ \dots \circ h)(a) \in \mathrm{Im}(y_0 - f) \,\forall k \leq z$ and $h^z(a) = a$. For example: $2 \overset{h}{\mapsto} 4 \overset{h}{\mapsto} 6 \overset{h}{\mapsto} 0 \overset{h}{\mapsto} 2$.
Proposition 7: Assume $Y = f(X) + N, N \perp\!\!\!\perp X$ and $n(l) \neq 0, p(k) \neq 0 \,\forall k, l$. Assume further that the model is reversible with a non-injective g.

- If h has at least one cycle, $\#\mathrm{Im}f - \#\text{cycles} + 1$ parameters of n are determined by the others.
- If h has no cycles, $\#\mathrm{Im}f$ parameters of n are determined by the others.

Proof: Assume h has a cycle of length r: $n_1 \overset{h}{\mapsto} n_2 \overset{h}{\mapsto} \dots \overset{h}{\mapsto} n_r \overset{h}{\mapsto} n_1$ (here, $y_0 - n_1, \dots, y_0 - n_r \in \mathrm{Im}f$), then $\frac{q(y_0)}{q(y_1)} = 1$ because $\frac{q(y_0)^r}{q(y_1)^r} = \frac{n(n_1)}{n(n_2)} \cdot \frac{n(n_2)}{n(n_3)} \dots \frac{n(n_r)}{n(n_1)} = \frac{n(n_1)}{n(n_1)} = 1$ and $n\big(y_0 - f(x)\big) = n\big(y_1 - f(x)\big) \,\forall x$, that is $n(n_1) = n(n_2) = \dots = n(n_r)$. Thus we get $r - 1$ equality constraints for each cycle of length r. For any (additional) non-cyclic structure of length r: $n_1 \mapsto n_2 \mapsto \dots \mapsto n_r$ and $n_r \notin \mathrm{Im}(y_0 - f)$ (here, $y_0 - n_1, \dots, y_0 - n_{r-1} \in \mathrm{Im}f$), we have $n(n_1) = \dots = n(n_r)$ and thus $r - 1$ equality constraints. Together with the normalization these are

$\#\mathrm{Im}f - \#\text{cycles} + 1$ constraints.

If h has no cycle, we have $\#\mathrm{Im}f - 1$ independent equations plus the sum constraint. E.g.: $\frac{n(2)}{n(4)} = \frac{n(4)}{n(6)} = \frac{n(3)}{n(5)}$ implies $n(4) = n(6)\frac{n(3)}{n(5)}$ and $n(2) = \frac{n(4)^2}{n(6)}$. Further,

$$\frac{n\big(y_0 - f(x)\big)}{n\big(y_1 - f(x)\big)} = \frac{q(y_0)}{q(y_1)} = \frac{\sum_{\tilde{x}} p(\tilde{x}) n(y_0 - f(\tilde{x})}{\sum_{\tilde{x}} p(\tilde{x}) n(y_1 - f(\tilde{x}))}$$

introduces a functional relationship between p and n. ■

Note that if $\tilde{m}$ does not have any divisors, there are no cycles and thus $\#\mathrm{Im}f$ parameters of n are determined. We have the following corollary

Corollary 8: In all cases the number of fixed parameters is lower bounded by $\lceil 1/2 \cdot \#\mathrm{Im}f \rceil + 1 \geq 2$.

3rd case: f is not injective.

Assume $f(x_0) = f(x_1)$. In a slight abuse of notation we write

$$g - g : \begin{array}{ccc} \mathbb{Z}/\tilde{m}\mathbb{Z} \times \mathbb{Z}/\tilde{m}\mathbb{Z} & \to & \mathbb{Z}/m\mathbb{Z} \\ (y, \tilde{y}) & \mapsto & g(y) - g(\tilde{y}) \end{array} .$$

Similar as above, we define

$$h_{x_0,x_1,g} : \begin{array}{ccc} \mathrm{Im}\big(x_0 - (g-g)\big) & \to & \mathbb{Z}/m\mathbb{Z} \\ x_0 - g(y) + g(\tilde{y}) & \mapsto & x_1 - g(y) + g(\tilde{y}) \end{array} .$$

We say that h has a cycle if there is a $z \in \mathbb{N}$, s.t. $h^k(a) = (h \circ \ldots \circ h)(a) \in \mathrm{Im}\big(x_0 - (g-g)\big)\ \forall k \leq z$ and $h^z(a) = a$.

Proposition 9: Assume $Y = f(X) + N$, $N \perp\!\!\!\perp X$, f is not injective and $n(l) \neq 0, p(k) \neq 0\ \forall k, l$. Assume further that the model is reversible for a function g.

- If h has at least one cycle, $\#\mathrm{Im}(g-g) - \#\text{cycles} + 1$ parameters of p are determined by the others.
- If h has no cycles, $\#\mathrm{Im}(g-g)$ parameters of p are determined by the others.

Proof: From (2) it follows that $\frac{p(x_0)}{p(x_1)} = \frac{\tilde{n}\big(x_0 - g(y)\big)}{\tilde{n}\big(x_1 - g(y)\big)} = \frac{p\big(x_0 - g(y) + g(\tilde{y})\big) \cdot n\Big(\tilde{y} - f\big(x_0 - g(y) + g(\tilde{y})\big)\Big)}{p\big(x_1 - g(y) + g(\tilde{y})\big) \cdot n\Big(\tilde{y} - f\big(x_1 - g(y) + g(\tilde{y})\big)\Big)}\ \forall y, \tilde{y}$. The rest follows analogously to the proof of Proposition 7. ■

If $(x_1 - x_0)$ does not divide m, there are no cycles and thus $\#\mathrm{Im}(g-g)$ parameters of p are determined.

Corollary 10: In all cases the number of fixed parameters is lower bounded by $\lceil 1/2 \cdot \#\mathrm{Im}(g-g) \rceil + 1 \geq 2$.

Remark 2: Note that some of the constraints described above depend on the backward function g. This introduces no problems because of the following reason: If we put any (prior) measure on the set of all possible parameters $p(0), p(1), \ldots, p(n-1)$ (or on $n(0), \ldots, n(m-1)$) that is absolutely continuous with respect to the Lebesgue measure, a single equality constraint reduces the set of possible parameters to a set of measure zero. There are only finitely many possibilities to choose the function g and thus even the union of all those parameter sets has measure zero.

VII. Conclusions and Future Work

We proposed a method that tries to infer the cause-effect relationship between two discrete random variables using the concept of ANMs. We proved that for generic choices the direction of a discrete ANM is identifiable in the population case and we developed an efficient algorithm that is able to apply the proposed inference principle to a finite amount of data from two variables; we also mentioned the limitations of p-values on real world data sets with changing data size.

Since it is known that χ^2 fails for small data sizes, changing the independence test for those cases may lead to an even higher performance of the algorithm. Further, our method can be generalized in different directions: (1) handling more than two variables is straightforward from a practical point of view, although one may have to introduce regularization to make the regression computationally feasible. Therefor, the work by Mooij et al. (2009) and the practical approach of combining constraint-based methods and ANMs by Tillman et al. (2009) may be helpful. (2) One should work on practically feasible extensions of ANMs and (3) it should be investigated how our procedure can be applied to the case, where one variable is discrete and the other continuous. Corresponding identifiability results remain to be shown. (4) We further believe that it is valuable to test the following principle for causal inference: One decides for $X \to Y$ not only if one finds an ANM in this direction and not the other, but also if the observed empirical distribution is *reasonably closer* to the set of distributions that allow for an ANM from X to Y than to the set of distributions that allow for an ANM from Y to X (e.g. the KL divergence to the subset $\mathrm{ANM}_{X \to Y}$ is smaller than to the subset $\mathrm{ANM}_{Y \to X}$, see section III-C). Clearly, the challenges of computing the distances to those sets of distributions and quantifying what *reasonably* closer means have to be solved.

We answered the theoretical question of identifiability, but in future work the concept of ANMs (like any proposed concept of causal inference) has to be addressed empirically: It should be tested on a large number of real world data sets, for which the ground truth is known. Our small collection of experiments, for example, only give a hint that ANMs may help for causal inference. It is still possible that exhaustive experiments show that the assumptions current methods for causal inference are based on are most often not met in nature. Nevertheless we regard our work as promising and hope that more fundamental and general principles for identifying causal relationships will be developed that cover ANMs as a special case.

Acknowledgement

The authors want to thank Fabian Gieringer for collaborating on section III-C and Joris Mooij, Kun Zhang and Stefan Harmeling for helpful comments.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd ed., 2002.

Armann, R., & Bülthoff, I. (2010). in preparation. `https://webdav.tuebingen.mpg.de/cause-effect/`.

Asuncion, A., & Newman, D. (2007). UCI machine learning repository. `http://archive.ics.uci.edu/ml/`.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10:417–451, 1954.

Czerniak, J., & Zarzycki, H. (2003). Application of rough sets in the presumptive diagnosis of urinary system diseases. In *Artifical Inteligence and Security in Computing Systems, 9th International Conference Proceedings*, 41–51. Kluwer Academic Publishers.

Heckerman, D., Meek, C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In C. Glymour, & G. Cooper, eds.,

Computation, Causation, and Discovery, 141–165. Cambridge, MA: MIT Press.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems (NIPS)*, 689–696. Vancouver, Canada: MIT Press.

Janzing, D., Peters, J., Mooij, J. M., & Schölkopf, B. (2009). Identifying confounders using additive noise models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 249–257. Corvallis, Oregon: AUAI Press.

Janzing, D., & Steudel, B. (2010). Justifying additive-noise-model based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, *17*, 189–212.

Kano, Y., & Shimizu, S. (2003). Causal inference using non-normality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, 261–270. Tokyo, Japan.

Mooij, J., Janzing, D., Peters, J., & Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 745–752. Montreal: Omnipress.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait. Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288).

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Peters, J., Janzing, D., Gretton, A., & Schölkopf, B. (2009). Detecting the direction of causal time series. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 801–808. Montreal: Omnipress.

Peters, J., Janzing, D., & Schölkopf, B. (2010). Identifying Cause and Effect on Discrete Data using Additive Noise Models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 9, 597–604.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag. (2nd edition MIT Press 2000).

Sun, X., Janzing, D., & Schölkopf, B. (2006). Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, 1–11. Fort Lauderdale, FL.

Sun, X., Janzing, D., & Schölkopf, B. (2008). Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomputing*, *71*, 1248–1256.

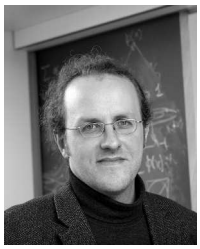
Tillman, R., Gretton, A. & Spirtes, P. (2009). Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2009.

Verma, T. & Pearl, J. (1991) Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc.

Zhang, K., & Hyvarinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 647–655. Corvallis, Oregon: AUAI Press.

Jonas Peters studied Mathematics and received a Master of Advanced Study and the UNWIN prize from Cambridge University (UK) in 2007 and a Diplom from the University of Heidelberg (Germany) in 2009. He wrote his diploma thesis about determining the direction of time series based on causality. For his PhD he is now working on causal inference problems at the Max Planck Institute for Biological Cybernetics in Tübingen. His studies were supported by the German National Academic Foundation, the DAAD and the Kurt-Hahn-Trust.

Dominik Janzing received a Diplom in Physics in 1995 and a PhD in Mathematics in 1998 from the University of Tübingen. From 1998-2006 he was a postdoc and senior scientist at the Computer Science department of the University of Karlsruhe (TH). Since 2007 he has been working as a senior scientist at the Max Planck Institute for Biological Cybernetics. There, he works on causal reasoning from statistical data. His approach uses complexity of conditional probability distributions. This idea is strongly influenced by his previous work on complexity of physical processes and the thermodynamics of information flow.



Bernhard Schölkopf received a doctorate in computer science from the Technical University Berlin. His thesis on Support Vector Learning won the annual dissertation prize of the German Association for Computer Science (GI). He has researched at AT&T Bell Labs, at GMD FIRST, Berlin, at the Australian National University, Canberra, and at Microsoft Research Cambridge (UK). In July 2001, he was appointed scientific member of the Max Planck Society. In 2006, he received the J. K. Aggarwal Prize of the International Association for Pattern Recognition. The ISI lists him as a highly cited researcher, and he is a board member of the NIPS foundation and of the International Machine Learning Society.