Identifying Cause and Effect on Discrete Data using Additive Noise Models

Jonas Peters
MPI for Biological Cybernetics
Spemannstr. 38
D-72076 Tübingen
jpeters@tuebingen.mpg.de

Dominik Janzing
MPI for Biological Cybernetics
Spemannstr. 38
D-72076 Tübingen
janzing@tuebingen.mpg.de

Bernhard Schölkopf
MPI for Biological Cybernetics
Spemannstr. 38
D-72076 Tübingen
bs@tuebingen.mpg.de

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. Recently, methods using additive noise models have been suggested to approach the case of 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. Whenever the joint distribution $\mathbf{P}^{(X,Y)}$ admits such a model in one direction, e.g. $Y = f(X) + N$, $N \perp\!\!\!\perp X$, it does not admit the reversed model $X = g(Y) + \tilde{N}$, $\tilde{N} \perp\!\!\!\perp Y$ as long as the model is chosen in a generic way. Based on these deliberations we propose an efficient new algorithm that is able to distinguish between cause and effect for a finite sample of discrete variables. We show that this algorithm works both on synthetic and real data sets.

1 Introduction

Inferring causal relations by analyzing statistical dependences among observed random variables is a challenging task if no controlled randomized experiments are available. So-called constraint-based approaches to causal discovery (Pearl, 2000; Spirtes et al., 1993) select among all directed acyclic graphs (DAGs) those that satisfy the Markov condition and the faithfulness assumption, i.e., those for which the observed independences are imposed by the structure rather than being a result of specific choices of parameters of the Bayesian network. These approaches are unable to distinguish among causal DAGs that impose the same independences. In particular, it is impossible to distinguish $X \to Y$ from $Y \to X$.

More recently, several methods have been suggested that use not only conditional independences, but also more sophisticated properties of the joint distribution. For simplicity, we explain the ideas for the two variable setting, a particularly challenging case. Kano & Shimizu (2003) and Shimizu et al. (2006) 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 from X to Y. Apart from trivial cases, $P(X, Y)$ can only admit such a model from X to Y and from Y to X in the bivariate Gaussian case, which leads to the following way of distinguishing between cause and effect: Whenever such an additive noise model exists in one direction but not in the other, we prefer the former based on Occam's Razor and infer it to be the causal direction. Janzing & Steudel (2009) give further theoretical support for this principle of causal inference using the concept of Kolmogorov complexity and Peters et al. (2009a) use this principle to detect whether a sample of a time series has been reversed. Hoyer et al. (2009) generalize the method to non-linear functions f and Zhang & Hyvarinen (2009) augment the model by applying a non-linear function g to the rhs of eq. (1). They still obtain identifiability for generic cases. All these proposals, however, were only designed for real-valued 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

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

norms. Sun et al. (2006); Janzing et al. (2009b) 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 additive noise model in eq. (1) to the discrete case by assuming both X and Y take values in $\mathbb{Z}$ (the support may be finite, though) and adopt the causal inference method from above: If there is an additive noise model from X to Y, but not vice versa, we infer that X is causing Y. Note that we can apply this principle even to the mixed case (one continuous and one discrete variable) by distretizing the continuous variable, assuming that this does not destroy the model structure.

Our causal inference method is sensible only if there are not many instances, for which there is an additive noise models in both directions. If all additive noise models from X to Y also allow an additive noise model from Y to X, for example, we could not draw any causal conclusions at all. We show that *reversible* cases are very rare and thereby answer this theoretical question.

For a practical causal inference method we need to test whether the data admits an additive noise model. In principle we thus have to check all possible functions and test whether they result in independent residuals. This is highly intractable since the function space is too large. In this work we propose an efficient heuristic procedure that proved to work very well in practice.

In section 2 we repeat the concept of additive noise models and show the corresponding identifiability results for generic cases in section 3. In section 4 we introduce an efficient algorithm for causal inference on finite data, for which we show experimental results in section 5. We conclude in section 6.

2 Additive Noise Models for Discrete Variables

For simplicity we introduce the following notation: $p_X(x) = \mathbf{P}(X = x)$, $p_Y(y) = \mathbf{P}(Y = y)$, $n(l) = \mathbf{P}(N = l)$, $\tilde{n}(k) = \mathbf{P}(\tilde{N} = k)$ and $\operatorname{supp} X$ is defined as the set of all values that X takes with probability larger than 0: $\operatorname{supp} X := \{k \,|\, p_X(k) > 0\}$.

Assume that X and Y take 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

$$Y = f(X) + N\,, \quad N \perp\!\!\!\perp X$$

where $f : \mathbb{Z} \to \mathbb{Z}$ is an arbitrary function and N a noise variable that takes integers as values, too.

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 (including a function g and some noise $\tilde{N}$) from Y to X, i.e. if it satisfies an ANM in both directions.

As it has been proposed for the continuous case we apply the following causal principle throughout the remainder of this article:

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

3 Identifiability

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 $p_X(0) = 1, n(0) = 1$ and $f(0) = 2$ immediately show that there are joint distributions allowing ANMs in both directions, meaning $F \cap B \neq \emptyset$. Our method, however, is only useful

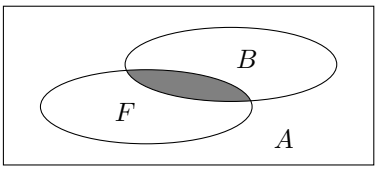

Figure 1: How large is $F \cap B$?

if the intersection is not too large. We therefore identify $F \cap B$ and show that it is indeed a very small set. If we are unlucky and the data generating process we consider happens to be in $F \cap B$, our method does not give wrong results, but answers "I do not know the answer". In all other situations the method identifies the correct direction given that we observe enough data.

3.1 X or Y has finite support

First we assume that either the support of X or the support of Y is finite. This already covers the situation in 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 $X = 5$: 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 $p_X(a_i) = \frac{1}{36}, p_X(b_i) = \frac{2}{36}$ for $i = 1, \ldots, 4$ and $p_X(a_i) = \frac{2}{36}, p_X(b_i) = \frac{4}{36}$ for $i = 5, \ldots, 8$.

The proofs of the following theorems are provided in (Peters et al., 2009b).

Theorem 1 *An additive noise model $X \to Y$ is reversible $\iff$ there exists a disjoint decomposition $\bigcup_{i=0}^{l} C_i = \operatorname{supp} X$, such that*

- *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: $f \,|_{C_i} \equiv c_i \; \forall i$.

- *The probability distributions on the C_is are shifted and scaled versions of each other with the same shift constant as above: For $x \in C_i$ the following equation holds*

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

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

By symmetry such a decomposition must exist for $\operatorname{supp} Y$, too. We are now given a full characterization of all cases that allow an ANM in both directions. Since already each condition by itself is very restrictive, all of them together describe a very small class of models: in almost all cases the direction of the model is identifiable. In Figure 3 all a_i belong to C_0, all b_j to C_1 and $d_1 = 1$. The main point of the general proof is based on the asymmetric effects of the "corners" of the joint distribution.

3.2 X and Y have infinite support

Theorem 2 *Consider an additive noise model $X \to Y$ where both X and Y have infinite support. We distinguish between two cases*

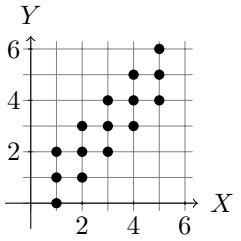

Figure 2: This joint distribution satisfies an additive noise model only from X to Y.

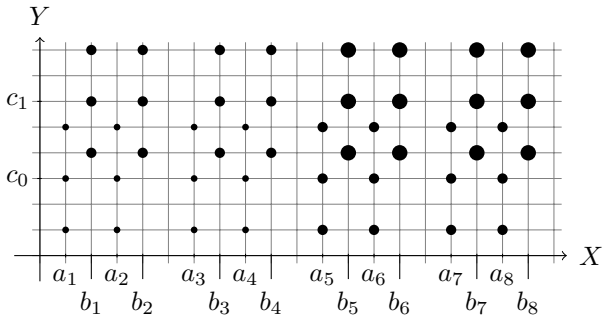

Figure 3: Choosing the parameters carefully leads to a reversible ANM.

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

2. **N has entire $\mathbb{Z}$ as support:** $\mathbf{P}(N = k) > 0 \; \forall k \in \mathbb{Z}$. *Suppose X and Y are not independent and there is an ANM $X \to Y$ and $Y \to X$. If f, the distribution of N and all $p_X(k)$ for all $k \geq m$ for any $m \in \mathbb{Z}$ are known, then all other values $p_X(k)$ for $k < m$ are determined. That means even only 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 we are in choosing a distribution $\mathbf{P}^X$ that yields a reversible ANM, for a given function f and noise N.

4 Practical Method for Causal Inference

Based on our theoretical findings in section 3 we propose the following method for causal inference (see Hoyer et al. (2009) for the continuous case):

(1) Given are iid data from (X, Y).

(2) Regression of the model $Y = f(X) + N$ leads to residuals $\hat{N}$,
regression of the model $X = g(Y) + \tilde{N}$ leads to residuals $\hat{\tilde{N}}$.

(3) If $\hat{N} \perp\!\!\!\perp X, \hat{\tilde{N}} \not\perp\!\!\!\perp Y$ infer *"X is causing Y"*,
if $\hat{N} \not\perp\!\!\!\perp X, \hat{\tilde{N}} \perp\!\!\!\perp Y$ infer *"Y is causing X"*,
if $\hat{N} \not\perp\!\!\!\perp X, \hat{\tilde{N}} \not\perp\!\!\!\perp Y$ infer *"I do not know (bad model fit)"* and
if $\hat{N} \perp\!\!\!\perp X, \hat{\tilde{N}} \perp\!\!\!\perp Y$ infer *"I do not know (both directions possible)"*.

We have shown before that the last condition will almost never occur. The procedure requires discrete methods for regression and independence testing and we now discuss our choices.

4.1 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 at least in principle not necessary in the discrete case. Since we may observe many different values of Y for one specific X value there is no risk in overfitting.

Minimizing a *loss function* like an ℓ_p error is not 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 Mooij et al. (2009) for the continuous case). Therefore we consider a dependence measure (DM) between residuals and regressor as loss function, which we denote by $\mathrm{DM}(\hat{N}, X)$.

Since we require $n(0) \geq n(k)$ for all $k \neq 0$, it is with high probability sufficient to regard only the values between $\min Y$ and $\max Y$ as possible values for f. If there are too few samples with $X = x_j$ and the value $f(x_j)$ is not included in $\mathcal{Y} := \{\min Y, \min Y + 1, \ldots, \max Y\}$ we may not find the true function f, but the few "wrong" residuals do not have an impact on the independence.
The search space, however, is still very large. In principle we have to try all of those functions and compare the corresponding values of the loss function. This is not tractable, of course: If there are 20 observed X values and $\#\mathcal{Y} = \max Y - \min Y + 1 = 16$, there are $16^{20} = 2^{80}$ possible functions. We propose the following 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 until convergence 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
$f^{(0)}(x_i) := \mathrm{argmax}_y \hat{\mathbf{P}}(X = x_i, Y = y)$
repeat
  $j = j + 1$
  for i in a random ordering **do**
    $f^{(j)}(x_i) := \mathrm{argmin}_y \mathrm{DM}\big(X, Y - f^{(j-1)}_{x_i \mapsto y}(X)\big)$
  end for
until residuals $\hat{N} := Y - f^{(j)}(X)$ are independent of X **or** $f^{(j)}$ does not change anymore.

Here, $f^{(j-1)}_{x_i \mapsto y}$ denotes the current version of $f^{(j-1)}$ but $f(x_i)$ changed to be y. If the argmax in the initialization step is not unique we take the largest possible y. If the argmin in the iteration step is not unique we take the y value that is closest to the old $f^{(j-1)}(x_i)$. The iteration can even be accelerated if we consider the five y values that give the largest $\hat{\mathbf{P}}(X = x_i, Y = y)$ instead of all possible values $\{\min Y, \ldots, \max Y\}$.

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 is known to converge 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.

The code for the proposed method is available at the author's homepage.

4.2 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 Person's χ^2 test (e.g. Lehmann & Romano (2005)), 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. For very few samples, however, this approximation and therefore the test will usually fail. It has been suggested (e.g. RProject (2009)) that instead of a χ^2 test, Fisher's exact test (Lehmann & Romano, 2005) could be used if not more than 80% of the expected counts are larger than 5 ("Cochran's condition"). For a dependence measure DM we simply use the 1 minus the p-value of the independence test or the test statistic if the p-value is too small (in a computer system the p-value is sometimes regarded to be zero).

5 Experiments

5.1 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 3 that only very few examples allow a reversible ANM. Data set 1 supports this theoretical result. We simulate a large amount of data from many randomly chosen models. All models that allow an ANM in both directions satisfy the conditions of Theorem 1 (without exception). Data set 2 shows how well our method performs for models that are close to non-identifiable and data set 3 investigates empirically the run-time performance of our regression method and compares it with a brute-force search.

Data set 1 (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 (n, p), (4) X geometric with parameter p, (5) X hypergeometric with (M, K, N), (6) X negative binomial with (n, p) or (7) X Poisson with parameter λ. The parameters of these distributions, the noise distribution (with values between -5 and 5) and the function (with values between -7 and 7) are also randomly chosen. We then consider 1000 different models. For each model we sample 1000 data points and apply our algorithm with $\alpha = 0.05$.

The results given in Table 1 show that the methods 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 which are instances of the counter examples from above. This experiment further supports our proposition that the model is identifiable in the generic case.

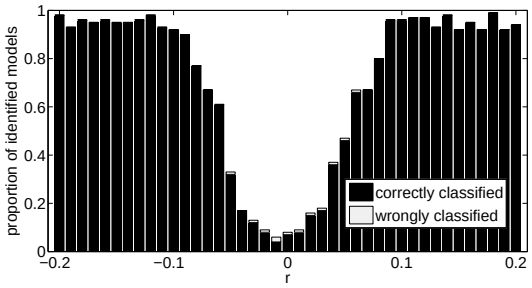

Figure 4: Data set 2. Proportion of correct and false results of the algorithm depending on the distribution of N. The model is not identifiable for $r = 0$. If r differs significantly from 0 almost all decisions are correct.

Data set 2 (close to non-identifiable).
For this data set we sample from the model $Y = f(X) + N$ with

$$n(-2) = 0.2,\ n(0) = 0.5,\ n(2) = 0.3 \quad \text{and}$$
$$f(-3) = f(1) = 1,\ f(-1) = f(3) = 2.$$

Depending on parameter r we sample X from

$$p_X(-3) = 0.1 + r/2,\ p_X(-1) = 0.3 - r/2,$$
$$p_X(1) = 0.15 - r/2,\ p_X(3) = 0.45 + r/2.$$

For each value of the parameter r ranging between $-0.2 \leq r \leq 0.2$ we use 100 different samples, each of which has the size 400.

In Theorem 1 we proved that the ANM is reversible if and only if $r = 0$. Figure 4 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 $|r| \geq 0.15$. The number of such cases can be reduced by decreasing α, but would lead to some more wrongly accepted backward models, too.

Data set 3 (fast regression).
The space of all functions from the domain of X to the domain of Y is growing very quickly in their sizes: If $\#\operatorname{supp} X = m$ and $\#\mathcal{Y} = \#\{\min Y, \ldots, \max Y\} = \tilde{m}$ then the proposed search space $\{f : \operatorname{supp} X \to \mathcal{Y}\}$ has $\tilde{m}^m$ elements. It is clear that it is infeasible to

# samples	correct dir.	wrong dir.	"both dir. poss."	"bad fit in both dir."
total	89.9%	0%	5.3%	4.8%
non-overlapping noise	-	-	3.0%	-
f constant	-	-	2.3%	-

Table 1: Data Set 1. The algorithm identifies the true causal direction in almost all cases. All models that were classified as reversible are either instances, where the noise does not "overlap" (i.e. $f(x) + \operatorname{supp} N$ are disjoint) or where f is constant. For the remaining models the algorithm mistakes the residuals as being dependent in 4.8% of the cases, which corresponds to the test level.

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 the algorithm we proposed in section 4.1 checks in order to find the true function (which it always did).

We simulated from the model $Y = \text{round}(0.5 \cdot X^2) + N$ for two different noise distributions:

$$n_1(-2) = n_1(2) = 0.05,$$
$$n_1(-1) = n_1(0) = n_1(1) = 0.3$$

and

$$n_2(-3) = n_2(3) = 0.05,$$
$$n_2(-2) = n_2(-1) = ... = n_2(2) = 0.18$$

Each time we simulated a uniformly distributed X with i values between $-\frac{i-1}{2}$ and $\frac{i-1}{2}$ for $i = 3, 5, 7, \ldots, 19$. 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 104 ± 33 functions before termination. The full results are shown in Figure 5.

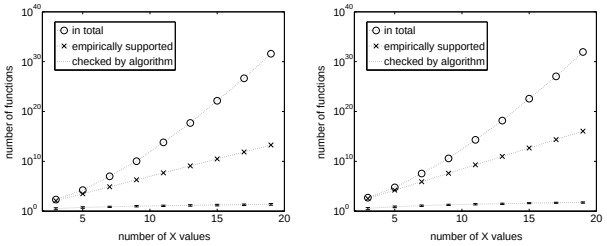

Figure 5: Data set 3. The size of the whole function space, the number of all functions with empirical support and the number of functions checked by our algorithm is shown for N_1 (left) and N_2 (right). An extensive search would be intractable in these cases. The algorithm we propose is very efficient and still finds the correct function for all data sets.

5.2 Real Data.

Cyclic Constraints
Note that the assumption of an ANM is unrealistic for many real world data sets: If Y can take only 2 values, for example, there is only little chance of fitting an ANM from X to Y. It is possible to extend the ANM to random variables that take values in a cyclic domain: $Y = f(X) + N$ then means Y and N take values in $\mathbb{Z}/m\mathbb{Z}$ and the $+$ should be interpreted as $+$ with $\bmod m$. This way we can model not only cyclic variables, but also variables that take categorical values (i.e. values in a structureless set): Therefore impose any cyclic structure on the values and use the additive noise $\mathbf{P}(N = 0) = p, \mathbf{P}(N = l) = (1 - p)/(m - 1)$ for $l \neq 0$.

To fit an ANM it does not make a difference if we assume the regressor X to be cyclic or not. Thus we only have to say whether we model the target variable as being cyclic (*cyclic constraint*) or not (*integer constraint*). Corresponding identifiable results (as in section 3) still hold. More details (including proofs) regarding cyclic constraints can be found in (Peters et al., 2009b). This model extension leads to more predictive power and thus to a better performance on real world data sets. To the best of our knowledge this is the first method that can take cyclic random variables into account.

Data set 4 (abalone).
We 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 (all considered in mm) and have $70, 57$ and 28 different values, respectively. Compared to the number of samples

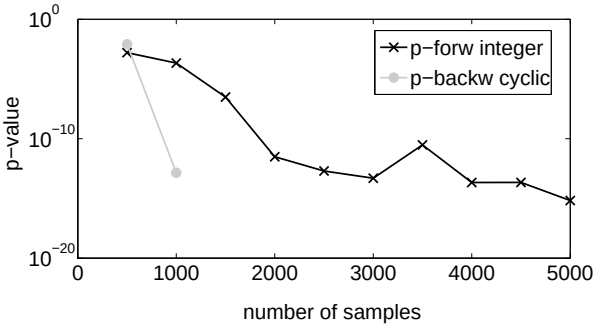

Figure 6: Data Set 5. The plots show p-values of forward and backward direction depending on the number of samples we included (no point means $p = 0$). The p-forward decrease much more slowly than p-backward.

(1000) 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.

Clearly, the Y variables do not have a cyclic structure. For the sex variable X we try both cyclic and non-cyclic constraints. Our method is able to identify all 3 directions correctly (see Table 2). We used $\alpha = 5\%$ and the first 1000 samples of the data set.

Data set 5 (temperature).
As Sun et al. (2006) we also applied our method to a data set consisting of 9162 daily values of temperature measured in Furtwangen (Germany) (Janzing, 2009) using the variables temperature (T, in $^\circ C$) and month (M). Clearly M inherits a cyclic structure, whereas T does not. Since the position of the earth relatively to the sun is causing the temperature, we take $M \to T$ as the ground truth. Here, we aggregate states and use months instead of days. This is done in order to meet Cochran's condition and get reliable results from the independence test (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 ($p\text{-value}_{M\to T} = 2\cdot 10^{-4}$, $p\text{-value}_{T\to M} = 1\cdot 10^{-13}$). But Figure 6 shows that the $p\text{-values}_{M\to T}$ are decreasing much more slowly than $p\text{-values}_{T\to M}$. Using other criteria than simple p-values we still may prefer the correct direction.

For both data sets the method proposed by (Janzing et al., 2009b) does not propose a causal direction because the difference in likelihoods is considered to be insignificant.

6 Conclusions and Future Work

We proposed a method that is able to infer the cause-effect relationship between two discrete random variables. We showed 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 infer the causal direction for a finite amount of data.

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. (2) It should further be investigated how our procedure can be applied (without discretization) to the case, where one variable is discrete and the other continuous. Corresponding identifiability results remain to be shown. (3) Since discrete data often originates from continuous data that have been measured and rounded, it may be useful to include models of the form $Y = g(f(X) + N)$ with g a thresholding function into the method. (4) For the continuous case Janzing et al. (2009a) try to identify the existence of a hidden common cause. A corresponding method for the discrete case could be used in order to distinguish between $X \to Y$ and $U \to X$, $U \to Y$ with unobserved U.

In future work ANMs should be tested on a large number of real world data sets with known ground truth in order to support (or disprove) ANMs as a principle in causal inference. We further hope that more general principles for identifying causal relationships will be developed that cover ANMs as a special case. Nevertheless we regard our work as a small step towards understanding the difference between cause and effect.

Acknowledgements

We thank Joris Mooij and the anonymous reviewers for their helpful comments.

References

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

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 D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou, eds., *Proceedings of*

	p-value$_{X\to Y}$	estimated function	p-value$_{Y\to X}$ (non-cyclic)	p-value$_{Y\to X}$ (cyclic)
Y_1	0.17	$0\mapsto 39, 1\mapsto 51, 2\mapsto 53$	$3\cdot 10^{-14}$	$3\cdot 10^{-2}$
Y_2	0.19	$0\mapsto 30, 1\mapsto 41, 2\mapsto 43$	$2\cdot 10^{-14}$	$4\cdot 10^{-3}$
Y_3	0.05	$0\mapsto 10, 1\mapsto 14, 2\mapsto 15$	0	$1\cdot 10^{-8}$

Table 2: Data Set 4. The algorithm identifies the true causal direction in all 3 cases. Here, we assumed a non-cyclic structure on Y and tried both cyclic and non-cyclic for X.

the conference Neural Information Processing Systems (NIPS) 2008, 689–696. Vancouver, Canada: MIT Press.

Janzing, B. (2009). Daily temperature (cause-effect pairs). `https://webdav.tuebingen.mpg.de/cause-effect/`.

Janzing, D., Peters, J., Mooij, J., & Schölkopf, B. (2009a). Identifying confounders using additive noise models. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*.

Janzing, D., & Steudel, B. (2009). Justifying additive-noise-model based causal discovery via algorithmic information theory. `http://arxiv.org/abs/0910.1691`.

Janzing, D., Sun, X., & Schölkopf, B. (2009b). Distinguishing cause and effect via second order exponential models. `http://arxiv.org/abs/0910.5561`.

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

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

Mooij, J., Janzing, D., Peters, J., & Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference. In L. Bottou, & M. Littman, eds., *Proceedings of the 26th International Conference on Machine Learning*, 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. (2009a). Detecting the direction of causal time series. In L. Bottou, & M. Littman, eds., *Proceedings of the 26th International Conference on Machine Learning*, 801–808. Montreal: Omnipress.

Peters, J., Janzing, D., & Schölkopf, B. (2009b). Causal inference on discrete data using additive noise models. `http://arxiv.org/abs/0911.0280v1`.

RProject (2009). The r project for statistical computing, see `?fisher.test`. Website, 15.1.2009, 1:07pm. Downloadable from `http://www.r-project.org/`.

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 (Lecture notes in statistics)*. New York, NY: Springer-Verlag.

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.

Zhang, K., & Hyvarinen, A. (2009). On the identifiability of the post-nonlinear causal model. *In: Proceedings of the 25th Annual Conference on Uncertainty in Artifical Intelligence*.