Causal Discovery by Randomness Test

S. Prestwich and **S. A. Tarim** and **I. Ozkan**
Insight Centre for Data Analytics
Department of Computer Science
University College Cork, Ireland

Abstract

Probabilistic methods for causal discovery are based on the detection of patterns of correlation between variables. They are based on statistical theory and have revolutionised the study of causality. However, when correlation itself is unreliable, so are probabilistic methods: nonsense correlations can lead to spurious causal links, while nonmonotonic functional relationships between variables can prevent the detection of causal links. We describe a new heuristic method for inferring causality between two continuous or integer variables, based on a nonparametric randomness test. We evaluate the accuracy of the method by comparing it to published algorithms on real and artificial datasets, and show that it largely avoids these false positives and negatives.

Introduction

Inferring cause-effect relationships between variables is of great importance in many areas, including medicine, sociology, bioinformatics, agriculture and most sciences. The standard scientific approach for determining such relationships is to design controlled experiments in which a few variables are manually changed and the results on other variables are observed. To many scientists this is the only acceptable method, but in applications where controlled experiments are expensive, unethical or infeasible we must try to infer causality from observed data only. In *causal discovery* we are given data consisting of observations of a number of variables, and our task is to infer causal relationships between the variables.

In recent years sophisticated probabilistic methods have been devised by Artificial Intelligence (AI) researchers for causal discovery in graphical form. Probabilistic methods are based on detecting patterns of conditional independence between variables, by applying a correlation test to some variables while conditioning on others. Algorithms such as IC (Pearl 2000), PC and FCI (Spirtes, Glymour, & Scheines 2000) have a sound statistical basis and have stimulated a great deal of interest and research. Non-probabilistic methods have also been reported in the AI literature. *Additive Noise Models* (Hoyer *et al.* 2009) can detect nonlinear causal relationships between two (or more) variables, by testing whether one variable y is a function of the other variable x, plus a noise term that is statistically independent of x. In this case there is usually no equivalent model with x and y interchanged, an asymmetry that is exploited to infer that x causes y ($x \rightarrow y$). *Information-Geometric Causal Inference* (Daniusis *et al.* 2010; Janzing *et al.* 2012) can also be applied to a pair of variables. It is based on the idea that the hypothesis $x \rightarrow y$ is only acceptable if the shortest description of $P_{x,y}$ (the joint probability distribution of x and y) is obtained from separate descriptions of P_x and $P_{y|x}$. The length of a description is defined in terms of its Kolmogorov complexity, which is uncomputable in principle but can be estimated. Outside AI there has been a great deal of research into causality. A survey of this vast literature is outside the scope of this paper, but see (Reiss 2015) for a recent overview.

A drawback of probabilistic methods is that nonmonotonic relationships between variables can be hard to detect by statistics such as Pearson's correlation coefficient. For example the correlation coefficient of $y = x^2$ with x sampled around 0 is close to 0, because the positive and negative correlations (for $x > 0$ and $x < 0$ respectively) tend to cancel out; similarly for periodic functions. Yet we might reasonably assume x to be a cause of y (assuming no other influences), because each x value has a unique y value but not vice-versa. As a more realistic example, we show below that electricity consumption depends nonmonotonically on temperature: house owners use heating when it is cold and air conditioning when it is hot, so consumption is least at intermediate temperatures.

Another drawback of probabilistic methods is their reliance on the *principle of the common cause* (PCC) (Reichenbach 1956): if two random variables x and y are probabilistically dependent then either $x \rightarrow y$, $y \rightarrow x$ or $x \leftarrow z \rightarrow y$ where z is a *confounding* (or *latent*) variable. But exceptions to the PCC have been reported. (Sober 2001) points out that two variables whose values increase monotonically (for example the Venetian sea level and British bread prices) will pass a correlation test though they are unrelated. These *nonsense* (or *illusory*) *correlations* were first described in (Yule 1926). (Hoover 2003) argues that this example relies on nonstationary data, for which correlation is an inappropriate measure of statistical association. (Spirtes, Glymour, & Scheines 2000) suggest that nonsense correlations might be caused by "remote unmeasured common causes". However, nonsense correlations can occur between time series that are both stationary and independent, as in this example

from (Reiss 2015):

$$x_t = \theta_x x_{t-1} + \epsilon_{x_t} \qquad y_t = \theta_y y_{t-1} + \epsilon_{y_t}$$

where $|\theta| < 1$ and the ϵ are independent and identically distributed random variables with zero mean. A dataset of such (x_t, y_t) pairs will exhibit many nonsense correlations (in about 30% of cases when $\theta = 0.75$). Preprocessing the data by differencing does not help because the mathematical form of the differenced series is identical to that of the original. Reiss also cites other ways in which nonsense correlations can occur, and concludes that there are no data preparation methods (such as differencing and detrending) that cure the problem, and no single causal inference algorithm that works well on all forms of data.

These observations do not contradict the work on probabilistic methods. They merely illustrate the known fact that those methods rest on assumptions that are violated by some forms of data (Spirtes, Glymour, & Scheines 2000). In this paper we describe a new causal discovery method designed to handle such data: it can discover nonmonotonic causal relationships between two continuous or integer variables without being misled by nonsense correlations. We describe our method, evaluate it on real and artificial data, and briefly discuss possible integrations with other methods.

A note on terminology based on (Salkind & Rasmussen 2007). Nonsense correlations are distinct from *spurious correlations*, which are accidental correlations that are not brought about by their claimed natural causes. They are artifacts of method and arise from factors such as sample selection bias or use of an inappropriate correlation coefficient. They can occur in *compositional data* such as proportions or percentages (Pearson 1897). The term *spurious correlation* may also refer to correlation caused by a confounding variable.

The algorithm

In this section we describe our new algorithm for inferring a causal relationship between two variables.

A causal discovery heuristic

Our algorithm is based on a simple heuristic:

> *Given two variables x and y, if y is everywhere a nonrandom function of x then infer $x \to y$.*

where *everywhere* means *for any subsequence of the sorted x values* (for simplicity we assume these are distinct). We can apply any convenient statistical test to decide whether or not a subsequence is a nonrandom function.

This is merely a heuristic and it is not hard to find examples in which it gives incorrect results. Firstly, suppose y is a monotonic function of x: then x will also appear to be a monotonic function of y. In this case our heuristic is unable to detect the direction of causality and will infer $x \leftrightarrow y$. Secondly, suppose y is a function of x with a plateau region (sometimes called *saturation* (Bunge 2009) Section 4.1.2). A plateau with added random noise looks like a random function, so y fails to be everywhere a nonrandom function of x and our heuristic will fail to identify the causal link $x \to y$.

Despite these limitations, we shall show that the heuristic leads to an interesting causal discovery algorithm that can match or outperform state-of-the-art methods. Before describing the algorithm we illustrate the motivation behind the heuristic.

Nonmonotonic causal relationships

As pointed out above, a nonmonotonic functional relationship can lead to a correlation coefficient close to zero, so that causal discovery algorithms based on correlations will not detect some causal links. In contrast, we propose to exploit such relationships.

First consider the $y = x^2$ example with added noise. Figure 1(a) shows a scatter plot of the data points. Would a human guess that $x \to y$ or $y \to x$? One way of deciding is to use a form of Occam's Razor (which is often invoked in causal inference research): choose whichever function looks most natural.

If $x \to y$ then we would expect y to be a reasonably well-behaved function of x, that is a function that does not look random. To test this hypothesis we could sort the data points by x then plot the y values as a function of $\text{rank}(x)$ using a line graph (where $\text{rank}(x)$ maps each x value to its rank in the dataset), giving the noisy but nonrandom-looking graph in Figure 1(b).

Similarly, to test the hypothesis $y \to x$ we could plot $\text{rank}(y)$ against x, obtaining the line graph shown in Figure 1(c). This is much less natural-looking and would represent a more complicated function. We conclude that y is a function of x and not vice-versa, hence that $x \to y$.

Next consider the dataset shown in Figure 1(d). Figure 1(e) plots y against $\text{rank}(x)$ and is nonrandom everywhere. Figure 1(f) plots $\text{rank}(y)$ against x: though much of the graph looks nonrandom, and the graph as a whole might pass a test of nonrandomness, there is a small random segment: it is not *everywhere* nonrandom. Hence we infer $x \to y$ and not $y \to x$.

Nonsense correlations

To illustrate the point that nonsense correlations can lead to spurious causal links, we created an artificial dataset with 50 variables and 200 observations. Each variable is independent of the others and its values are constructed by time series:

$$x_{i,1} = N(0, \sigma) \qquad x_{i,t} = x_{i,t-1} + N(0, \sigma)$$

Because the variables are independent and are not caused by any confounding variable the causal graph should be empty. However, applying the PC algorithm (Spirtes, Glymour, & Scheines 2000) with $\alpha = 0.05$ yields a causal graph that involves all 50 variables each with 1–4 links to other variables.

Figure 2(a) shows an example of two time series. Despite being independently generated they are negatively correlated, which misleads correlation-based methods. Figure 2(b) plots y against $\text{rank}(x)$ and 2(c) plots $\text{rank}(y)$ against x. Both look noisy but non-random: (b) has a general downward trend while (c) has two distinct segments. However, neither is *everywhere* nonrandom because each contains rather long subsequences that look random, so our

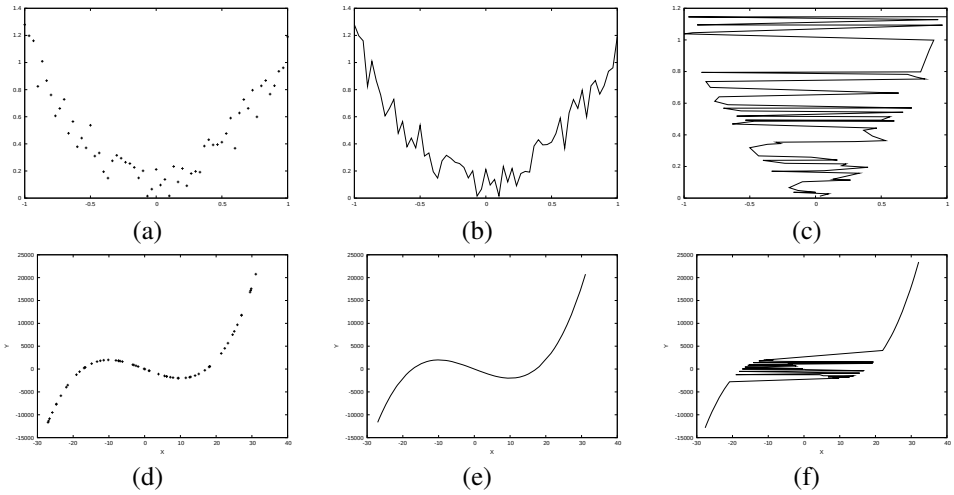


Figure 1: Inferring nonmonotonic causality between two variables

heuristic detects no causal link between the variables. For the 50-variable example our algorithm (described below) correctly finds no causal links.

The RCI algorithm

Whereas correlation is a symmetric statistic defined on two variables, we use an asymmetric statistic based on the well-known Wald-Wolfowitz runs test (Wald & Wolfowitz 1940). In the runs test we are given a binary sequence containing N_0 zeroes and N_1 ones, and we test the null hypothesis that the sequence was randomly generated, that is independently drawn from the same distribution. This is a nonparametric test as it does not rely on any assumptions regarding probability distributions. To test the hypothesis we count the number of *runs* R in the sequence, where a run consists of all zeroes or all ones. For example the sequence 0010111011 contains $R = 6$ runs (00, 1, 0, 111, 0 and 11). We also compute the expected number of runs

$$\bar{R} = \frac{2N_0N_1}{N_0 + N_1} + 1$$

and its standard deviation

$$S = \sqrt{\frac{(\bar{R} - 1)(\bar{R} - 2)}{N - 1}}$$

Finally we compute the test statistic

$$Z = \frac{\bar{R} - R}{S}$$

and reject the null hypothesis if $|Z|$ is greater than a threshold $\hat{Z}$ chosen to correspond to a significance level α (for example a value of 1.96 corresponds to an α of 5% and 2.58 to 1%); we shall denote the significance level by α. (The above formulae are based on a normal approximation and are only useful for a reasonably large number of samples such as $N > 30$, so for smaller N we compute $\bar{R}$ and S by brute force enumeration.)

The runs test can be used as a goodness-of-fit test: to test whether a curve $y = f(x)$ fits a dataset $\{(x_i, y_i) \,|\, i = 1, 2, \ldots\}$ derive a binary sequence b_i where $b_i = 1$ if $y_i > f(x_i)$ and 0 otherwise, then if the sequence passes the randomness test the curve is considered to be a good fit.

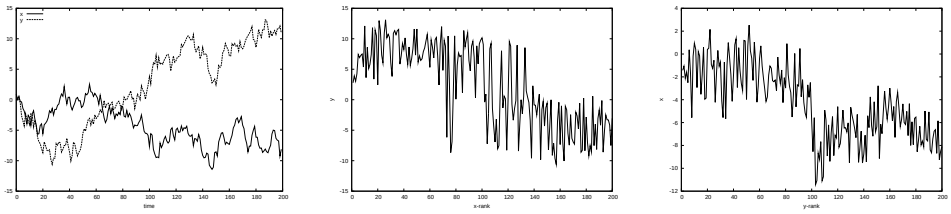
We use the runs test for causal inference as follows. Suppose we have a set S of n observations of two variables (x, y). To test whether y is everywhere a nonrandom function of x we sort the observations by x and extract the ordered list L of y values. For each sublist L' in a selected set $\mathrm{s}(L)$ of sublists of L, we compute the Z statistic for a goodness-of-fit test to the flat line $y = \mu$ where $\mu = \mathrm{mean}(L')$. We denote the least such Z by $Z_x(S)$; similarly for $Z_y(S)$. The set $\mathrm{s}(L)$ is computed as follows:

$\mathrm{s}(L) =$
  $m = \mathrm{length}(L)$
  if $m < \ell$
    return $\emptyset$
  else
    return $\{L\} \cup \mathrm{s}(L_{1\ldots\lceil 2m/3\rceil}) \cup \mathrm{s}(L_{\lfloor m/3\rfloor\ldots m})$

We do not consider sequences shorter than some minimum length ℓ, unless $n < \ell$ in which case we use $\{L\}$ instead of $\mathrm{s}(L)$. Now if $Z_x > \hat{Z}$ we infer $x \to y$, and if $Z_y > \hat{Z}$ we infer $y \to x$.

We call this algorithm *Randomness-based Causal Inference* (RCI). The time complexity of RCI on n observations of v variables is $O(v^2 n \log n)$: each pair of variables are tested for causality, and the time for each pair is dominated by the need to sort pairs of values. However, this assumes that we require a full causal graph: if we are only interested in the causes and/or effects of a single variable then we need only consider $v - 1$ possible causal links, and the complexity is reduced to $O(vn \log n)$. The algorithm has two parameters that a user must tune: ℓ and α (or $\hat{Z}$).

A further modification is required to handle a feature of many datasets: each x may have multiple y values and vice-versa, perhaps because of quantisation or rounding. For such



(a) (b) (c)

Figure 2: Time series example

data we simply take the median of the y values for each x value.

Experiments

We now apply RCI to several datasets and compare its results with those of other causal inference algorithms, including the PC algorithm implemented in the R software package (Kalisch *et al.* 2012). Unless stated otherwise we set the PC parameter $\alpha = 0.05$, and the RCI parameters $\ell = 50$ or $0.1n$ whichever is greater (to reduce plateau effects and avoid randomness tests on short subsequences) and $\hat{Z} = 2.58$ (hence $\alpha = 1\%$).

The Tübingen CEP benchmarks

Our algorithm works on pairs of variables so we start by testing it on bivariate datasets. A collection of these is maintained by the Max-Planck-Institute for Intelligent Systems at Tübingen (Mooij *et al.* 2014a) and an extension of a collection of datasets from a competition held in a causality workshop in 2008. The Cause Effect Pairs (CEP) collection is continually being updated and at the time of writing has 98 examples, but we use examples 1–88 so that we can compare our results with those in (Mooij *et al.* 2014b).

The CEP benchmarks were obtained from 31 datasets in a variety of domains: abalone measurements, census income, fuel consumption, geyser eruption, concrete properties, car traffic, ozone levels, UN statistics, stock returns, internet traffic, human face classification, sunspots, food and agriculture, light response, US country growth, milk protein, supply and demand for accommodation, environmental factors, climate and meteorology, and medicine. In each case there are two variables and the direction of causality (the *ground truth*) is self-evident.

Because the datasets are grouped into families with similar characteristics, the Tübingen group weighted the results so that each family's best possible marks summed to unity, giving a simple way of assigning an accuracy measure to the results of a causal inference method. (Mooij *et al.* 2014b) presented results using nine Additive Noise Models and three Information-Geometric Causal Inference methods. The results are plotted in graphs so we do not have exact figures, but we can estimate them by inspection: the highest accuracy was approximately 70–75% achieved by variant "ent-KDP" while most variants achieved below 65%.

The RCI results are shown in Table 1 which shows the values of Z_x and Z_y, and the inferred direction of causality. Correct inferences are highlighted by *. Under the accuracy measure we achieved 74.6%. As pointed out by (Mooij *et al.* 2014b) it is hard to know how significant these results are, and some of the choices made by RCI are probably not meaningful: for example instances 43–46 all have large positive causal strengths in both directions, while 81–83 have negative strengths in both directions. The results are encouraging but more experiments are required.

Electricity consumption

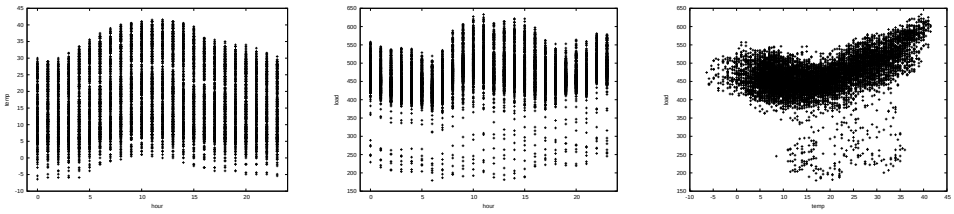
Next we take a dataset from the energy industry (the data is real but its source is confidential). The load (kW/hour) on a power station was measured hourly for 9504 hours, and the outside temperature was measured at the same times. So we have three variables: hour (integers in the range 0–23), temp (real values rounded to the nearest 0.5) and load (real values to four decimal places). Three bivariate scatterplots of the data are shown in Figure with the following x-y pairs: (a) hour vs temp, (b) hour vs load, and (c) temp vs load. The authors recently submitted these datasets to the Tübingen CEP collection as pairs 94–96.

We would like to know which variables cause which. By common sense we know that hour and temperature might influence load but not vice-versa, and that hour influences temperature. Each of these dependencies corresponds to a nonmonotonic function. The temperature peaks at approximately mid-day, the load appears to follow a slightly more complex pattern as the day proceeds, and the load increases as low and high temperatures: at low temperatures heating is used, while at high temperatures air conditioning is used (Taieb & Hyndman 2014).

RCI finds causal links hour→temp, hour→load and temp→load as expected. PC with $\alpha = 0.05$ finds undirected links between all three variables, while with smaller α it finds hour→load and temp→load but not hour→temp. We conjecture that the nonmonotonic nature of temperature with respect to hour confuses the correlation measure used by PC.

Communities and crime

Finally, from the UCI Machine Learning Repository (Lichman 2013) we obtained the Communities and Crime dataset (Redmond & Baveja 2002) which has 128 variables and 1994 observations. We ignore 25 of the variables which are



(a) (b) (c)

Figure 3: Three views of the electricity consumption data

textual, nominal or have missing values. The variable to be predicted in this dataset is ViolentCrimesPerPop (total number of violent crimes per 100K population).

As we are only interested in the causes of crime we use RCI to look for direct causal links between ViolentCrimesPerPop and all other variables. PC indicates that some variables are both causes and effects of ViolentCrimesPerPop because its CPDAG contains loops: for example a 3-loop involving PctIlleg, PctVacantBoarded and ViolentCrimesPerPop. So to compare the two methods we look at the variables they indicate are causes of crime, in the case of PC both direct and indirect (mediated by other variables). The PC causal graph is too large to display here but PC found 26 causes:

> householdsize, agePct12t21, agePct16t24, pctWInvInc, pctWPubAsst, NumUnderPov, PctBSorMore, PctUnemployed, PctOccupMgmtProf, MalePctDivorce, MalePctNevMarr, FemalePctDiv, TotalPctDiv, PersPerFam, PctFam2Par, PctKids2Par, PctYoungKids2Par, PctTeen2Par, NumIlleg, PctIlleg, NumImmig, PersPerOwnOccHous, PctHousOwnOcc, PctVacantBoarded, NumInShelters, NumStreet

whereas RCI found 56 causes:

> population, racePctBlack, racePctWhite, racePctAsian, agePct12t21, agePct12t29, agePct16t24, numbUrban, medIncome, pctWWage, pctWFarmSelf, pctWInvInc, pctWSocSec, pctWPubAsst, pctWRetire, medFamInc, perCapInc, blackPerCap, indianPerCap, OtherPerCap, NumUnderPov, PctNotHSGrad, PctBSorMore, PctOccupMgmtProf, MalePctDivorce, FemalePctDiv, TotalPctDiv, PctKids2Par, PctYoungKids2Par, PctTeen2Par, PctWorkMomYoungKids, NumIlleg, PctIlleg, NumImmig, PctImmigRecent, PctImmigRec5, PctImmigRec8, PctImmigRec10, PctRecentImmig, PctSpeakEnglOnly, PctNotSpeakEnglWell, PctLargHouseFam, PersPerOccupHous, PersPerOwnOccHous, PctPersDenseHous, PctHousLess3BR, PctHousOccup, PctHousOwnOcc, PctVacantBoarded, PctVacMore6Mos, MedRentPctHousInc, NumInShelters, PctForeignBorn, PctBornSameState, PctSameHouse85, PctUsePubTrans

It is reasonable to ask whether RCI is more sensitive or merely less discerning than PC, and whether one set of causes is more correct than the other. Both sets seem reasonable, but this is to be expected because all the variables were considered possible causes of crime by the researchers who collected the data.

One way of comparing the results is to analyse how closely they agree. We are using PC as a benchmark against which to evaluate RCI, so let us assume that PC has found the correct $K = 26$ causes out of $N = 102$ possibilities. Then of the $n = 56$ causes found by RCI $k = 20$ are correct. We can compute the probability that at least k of the n are correct by using the probability mass function of the hypergeometric distribution:

$$p(\geq k \text{ successes}) = \sum_{i=k}^{K} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$$

The probability that RCI finds at least 20 correct causes is $p = 0.0076$. Thus if our null hypothesis is that RCI chose its 56 causes randomly, we can reject the null hypothesis with significance 1%. Hence there is a high degree of agreement between the two methods on the causes of crime (though RCI is not necessarily correct to indicate 30 extra causes).

Conclusion

We described a new causal discovery algorithm called RCI and showed that it:

- can detect causal links that are undetected by a probabilistic algorithm;
- can avoid spurious causal links that are found by a probabilistic algorithm;
- performs well on bivariate datasets compared to additive-noise and information-geometric methods;
- performs well on multivariate datasets compared to a probabilistic algorithm.

It has a low worst-case time complexity, making it applicable to fairly large datasets. Applications for fast, approximate causal inference methods include causal feature selection (Guyon, Aliferis, & Elisseeff 2007).

RCI might be combined with other methods to yield a more powerful causal inference system. For example if RCI detects a bidirectional link caused by a monotonic function relating the two variables, this might be reduced to a unidirectional link by additive-noise or information-geometric methods. As another example, if we infer $a \leftrightarrow b \leftrightarrow c$ then we could invoke a rule described by Pearl to refine this to $a \rightarrow b \leftarrow c$.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. S. Armagan

Tarim is supported by the Scientific and Technological Research Council of Turkey (TUBITAK). Our research was aided by the availability of benchmarks in the UCI Machine Learning Repository (Lichman 2013) and the Cause Effect Pairs collection of (Mooij *et al.* 2014a). Thanks to Carlo Manna and Dalila Messedi for help with our research into causality.

References

Bunge, M. 2009. *Causality and Modern Science*. Transaction Publishers.

Daniusis, P.; Janzing, D.; Mooij, J. M.; Zscheischler, J.; Steudel, B.; Zhang, K.; and Schoelkopf, B. 2010. Inferring deterministic causal relations. In *26th Conference on Uncertainty in Artificial Intelligence*, 143–150.

Guyon, I.; Aliferis, C.; and Elisseeff, A. 2007. Causal feature selection. In Liu, H., and Motoda, H., eds., *Computational Methods of Feature Selection*. Chapman and Hall/CRC.

Hoover, K. D. 2003. Nonstationary time series, cointegration, and the principle of the common cause. *Brit. J. Phil. Sci.* 54:527–551.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schoelkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, 689–696.

Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniusis, P.; Steudel, B.; and Schoelkopf, B. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182-3:1–31.

Kalisch, M.; Maechler, M.; Colombo, D.; Maathuis, M. H.; and Buehlmann, P. 2012. Causal inference using graphical models with the R package. *Journal of Statistical Software* 47(11).

Lichman, M. 2013. UCI machine learning repository.

Mooij, J. M.; Janzing, D.; Zscheischler, J.; and Schoelkopf, B. 2014a. CauseEffectPairs repository http://webdav.tuebingen.mpg.de/causality/.

Mooij, J. M.; Peters, J.; Janzing, D.; Zscheischler, J.; and Schoelkopf, B. 2014b. Distinguishing cause from effect using observational data: Methods and benchmarks. Technical Report arXiv:1412.3773v1, Max-Planck-Institute for Intelligent Systems at Tuebingen.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearson, K. 1897. Mathematical contributions to the theory of evolution — on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 60:489–498.

Redmond, M. A., and Baveja, A. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141:660–678.

Reichenbach, H. 1956. *The Direction of Time*. Berkeley: University of California Press.

Reiss, J. 2015. *Causation, Evidence, and Inference*. Routledge.

Salkind, N. J., and Rasmussen, K. 2007. *Encyclopedia of Measurement and Statistics*. SAGE Publications Inc.

Sober, E. 2001. Venetian sea levels, British bread prices, and the principle of the common cause. *Brit. J. Phil. Sci.* 52:331–346.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction and Search*. MIT Press, Cambridge.

Taieb, S. B., and Hyndman, R. J. 2014. A gradient boosting approach to the kaggle load forecasting competition. *International Journal of Forecasting* 30(2):382–394.

Wald, A., and Wolfowitz, J. 1940. On a test whether two samples are from the same population. *Ann. Math. Statist.* 11:147–162.

Yule, G. 1926. Why do we sometimes get nonsense-correlations between time series? *Journal of the Royal Statistical Society* 89:1–64.

instance	truth	Z_x	Z_y	infer	
1	$x \to y$	0.1	3.9	$y \to x$	
2	$x \to y$	-0.1	0.2	$y \to x$	
3	$x \to y$	-0.2	1.5	$y \to x$	
4	$x \to y$	0.4	-1.1	$x \to y$	*
5	$x \to y$	5.0	10.1	$y \to x$	
6	$x \to y$	5.0	2.0	$x \to y$	*
7	$x \to y$	5.0	9.6	$y \to x$	
8	$x \to y$	5.0	6.6	$y \to x$	
9	$x \to y$	5.0	-0.6	$x \to y$	*
10	$x \to y$	5.0	-0.8	$x \to y$	*
11	$x \to y$	5.0	1.4	$x \to y$	*
12	$x \to y$	4.2	0.0	$x \to y$	*
13	$x \to y$	0.6	-2.2	$x \to y$	*
14	$x \to y$	2.4	0.7	$x \to y$	*
15	$x \to y$	-2.2	0.0	$y \to x$	
16	$x \to y$	-1.2	0.3	$y \to x$	
17	$x \to y$	4.5	-0.2	$x \to y$	*
18	$x \to y$	-0.3	1.8	$y \to x$	
19	$x \to y$	0.1	1.0	$y \to x$	
20	$x \to y$	0.6	2.2	$y \to x$	
21	$x \to y$	0.3	-3.1	$x \to y$	*
22	$x \to y$	-0.3	-0.5	$x \to y$	*
23	$x \to y$	1.1	0.3	$x \to y$	*
24	$x \to y$	-1.0	-0.1	$y \to x$	
25	$x \to y$	1.5	-1.7	$x \to y$	*
26	$x \to y$	1.4	-0.6	$x \to y$	*
27	$x \to y$	0.1	-0.9	$x \to y$	*
28	$x \to y$	0.5	-1.8	$x \to y$	*
29	$x \to y$	-0.5	-2.6	$x \to y$	*
30	$x \to y$	2.5	-0.6	$x \to y$	*
31	$x \to y$	-0.1	-1.8	$x \to y$	*
32	$x \to y$	3.3	-1.1	$x \to y$	*
33	$x \to y$	2.6	1.9	$x \to y$	*
34	$x \to y$	1.4	-2.0	$x \to y$	*
35	$x \to y$	1.4	1.1	$x \to y$	*
36	$x \to y$	2.6	3.3	$y \to x$	
37	$x \to y$	1.4	-1.6	$x \to y$	*
38	$x \to y$	0.5	-1.9	$x \to y$	*
39	$x \to y$	1.5	-2.3	$x \to y$	*
40	$x \to y$	3.2	2.6	$x \to y$	*
41	$x \to y$	0.8	-0.1	$x \to y$	*
42	$x \to y$	18.3	12.7	$x \to y$	*
43	$x \to y$	23.7	23.5	$x \to y$	*
44	$x \to y$	28.9	29.5	$y \to x$	
45	$x \to y$	19.3	20.7	$y \to x$	
46	$x \to y$	16.5	16.4	$x \to y$	*
47	$y \to x$	-0.7	0.0	$y \to x$	*
48	$y \to x$	-1.2	0.1	$y \to x$	*
49	$y \to x$	-1.1	-2.1	$x \to y$	
50	$y \to x$	-0.9	-0.9	$y \to x$	*
51	$y \to x$	-3.0	0.0	$y \to x$	*
56	$y \to x$	3.2	-1.4	$x \to y$	
57	$y \to x$	3.5	-1.3	$x \to y$	
58	$y \to x$	3.9	-1.3	$x \to y$	
59	$y \to x$	2.6	-1.4	$x \to y$	
60	$y \to x$	2.9	-1.6	$x \to y$	
61	$y \to x$	3.1	-1.2	$x \to y$	
62	$y \to x$	1.8	-1.1	$x \to y$	
63	$y \to x$	1.8	-0.8	$x \to y$	
64	$x \to y$	4.4	0.9	$x \to y$	*
65	$x \to y$	-0.9	-1.4	$x \to y$	*
66	$x \to y$	-1.8	-2.3	$x \to y$	*
67	$x \to y$	-2.3	-1.5	$y \to x$	
68	$y \to x$	0.0	2.1	$y \to x$	*
69	$y \to x$	2.8	6.0	$y \to x$	*
70	$x \to y$	3.5	0.0	$x \to y$	*
72	$x \to y$	-1.1	-2.2	$x \to y$	*
73	$y \to x$	-0.3	-2.0	$x \to y$	
74	$x \to y$	-2.7	5.9	$y \to x$	
75	$y \to x$	-0.7	-1.3	$x \to y$	
76	$x \to y$	5.2	4.1	$x \to y$	*
77	$y \to x$	15.9	-2.9	$x \to y$	
78	$x \to y$	0.2	-0.6	$x \to y$	*
79	$y \to x$	-0.6	-2.6	$x \to y$	
80	$y \to x$	-1.8	-1.1	$y \to x$	*
81	$x \to y$	-0.5	-0.7	$x \to y$	*
82	$x \to y$	-1.2	-1.1	$y \to x$	
83	$x \to y$	-0.3	-1.4	$x \to y$	*
84	$y \to x$	3.7	4.2	$y \to x$	*
85	$x \to y$	2.1	-0.3	$x \to y$	*
86	$x \to y$	5.2	3.0	$x \to y$	*
87	$x \to y$	8.9	-1.8	$x \to y$	*
88	$x \to y$	-1.2	-1.7	$x \to y$	*

Table 1: RCI results on CEP benchmarks