Alternative Approaches to Discovering Causality with Additive Noise Models

Kaleigh Clary and David Jensen

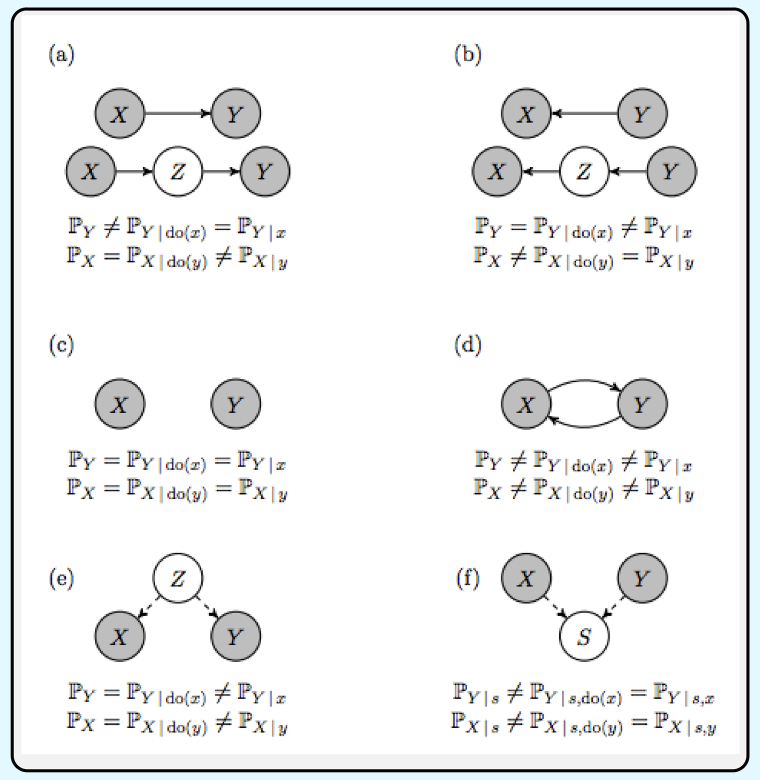
Knowledge Discovery Laboratory • College of Information and Computer Sciences • University of Massachusetts Amherst

1. Motivation

Bivariate causality is a branch of causality focused on inferring the causal direction between a pair of variables based on observational data.

These causal models assume some causal relationship exists between two variables *X* and *Y* as in (a) and (b), and there are no (d) feedback relationships or (e), (f) confounding variables *Z*.

Given only observational data from the distribution *P(X,Y)*, we wish to determine the direction of the causal edge between *X* and *Y*.



Reproduced from [1].

The accuracy of existing bivariate additive noise models is reported at 65-85% [1]. Many possible variations in the basic method have not been explored. We present results for bivariate ANM using alternate regression methods, using *k*-fold cross-validation to obtain residuals, and using data preprocessing to improve approximation of the regression function.

2. Background on additive noise models

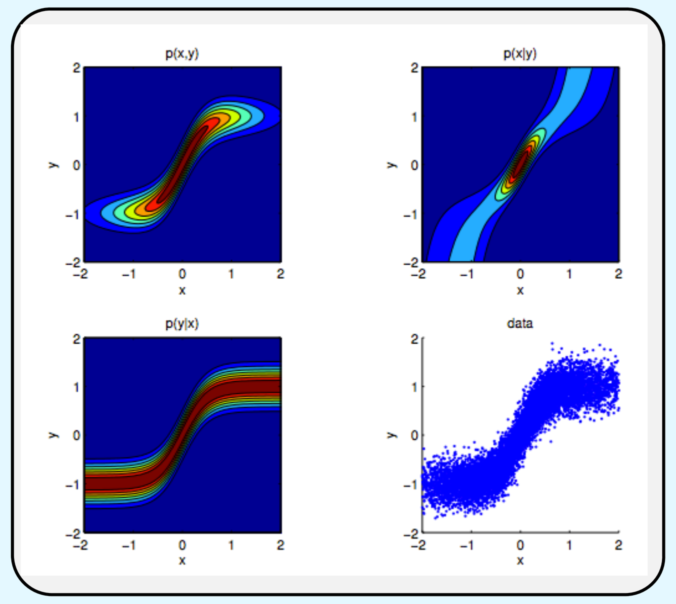
Bivariate additive noise models extend structural equation models which model effects as a function of their causes and latent noise variables.

ANMs model one variable as a function of the other and test for independence between that variable and the residuals of the regression function.

Two models are constructed, one in each causal direction, and the independence scores are directly compared for causal inference.

The contour lines of *P(Y|X)* change as X varies, but the contour lines of P(X|Y) are considerably more variable. This distribution **satisfies** an ANM *X* ➞ *Y*.

By assuming *a priori* that the distribution only satisfies an ANM in one direction, models in both directions can be directly compared.



Reproduced from [1].

3. Bivariate causal inference algorithm

Bivariate ANM algorithm for inferring directionality between two variables

For a given regression method and scoring function:

1. Construct a regression model in both causal directions.
2. Calculate residuals of both regression functions.

3, 4. Select the model with the least measured independence between the independent variable and residuals.

Algorithm 1 General procedure to decide whether $p(x,y)$ satisfies an additive noise model $X \to Y$ or $Y \to X$.

Input:

1. an i.i.d. sample $\mathcal{D}_N := \{(x_i, y_i)\}_{i=1}^N$ of X and Y ("training data");
2. an i.i.d. sample $\mathcal{D}'_N := \{(x'_i, y'_i)\}_{i=1}^N$ of X and Y ("test data").

Parameters:

1. Regression method
2. Score function $C : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ for measuring dependence

Output: one of $\{X \to Y, Y \to X, ?\}$.

1. Use the regression method to obtain estimates:
   (a) $\hat{f}_Y$ of the regression function $x \mapsto \mathbb{E}(Y \mid X = x)$,
   (b) $\hat{f}_X$ of the regression function $y \mapsto \mathbb{E}(X \mid Y = y)$
   using the training data $\mathcal{D}_N$;
2. Use the estimated regression functions to predict residuals:
   (a) $\hat{\boldsymbol{e}}'_Y := \boldsymbol{y}' - \hat{f}_Y(\boldsymbol{x}')$
   (b) $\hat{\boldsymbol{e}}'_X := \boldsymbol{x}' - \hat{f}_X(\boldsymbol{y}')$
   from the test data $\mathcal{D}'_N$.
3. Calculate the scores to measure independence of inputs and estimated residuals on the test data $\mathcal{D}'_N$:
   (a) $C_{X \to Y} := C(\boldsymbol{x}', \hat{\boldsymbol{e}}'_Y)$;
   (b) $C_{Y \to X} := C(\boldsymbol{y}', \hat{\boldsymbol{e}}'_X)$;
4. Output:

$$\begin{cases} X \to Y & \text{if } C_{X\to Y} < C_{Y \to X}, \\ Y \to X & \text{if } C_{X\to Y} > C_{Y \to X}, \\ ? & \text{if } C_{X\to Y} = C_{Y \to X}. \end{cases}$$

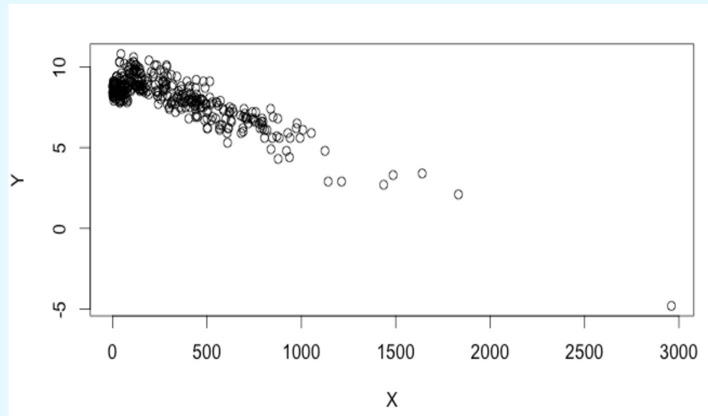
Reproduced from [1].

Scoring Methods:

- Approximation of shannon entropy
- Hilbert-Schmidt Independence Criterion (HSIC) statistic
- HSIC *p*-value
- HSIC statistic with fixed kernel
- Gaussian, or sum of variance of independent variable and residuals

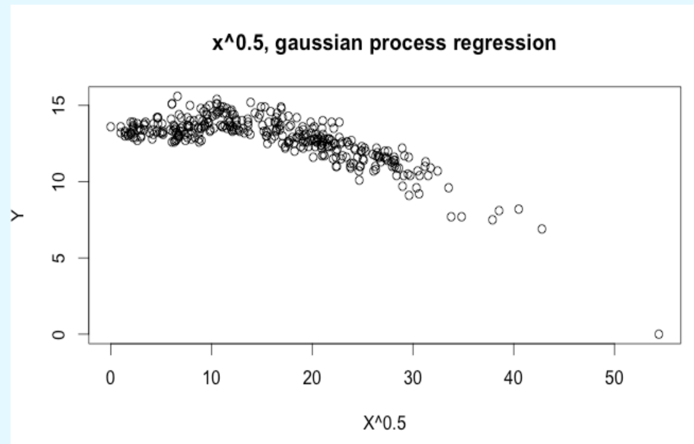
4. Background on Tukey's ladder of powers

In many pairs of the CEP dataset, data is highly concentrated in one are and sparse elsewhere.

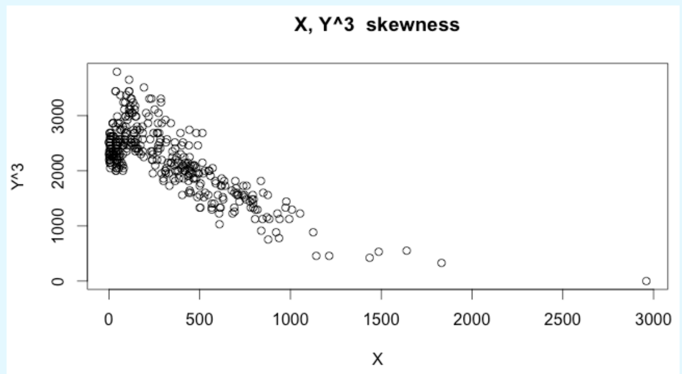


Interpolation or downsampling data will introduce additional noise and may bias results. Instead, we apply Tukey's ladder of powers to transform each pair dataset.

For a simple linear function Y = f(X) + ε Tukey's transformation applies a power λ to a variable as Y = f(X^λ) + ε. We set λ ϵ [-3, 3] and when λ = 0, we define X^λ = log(*X*).

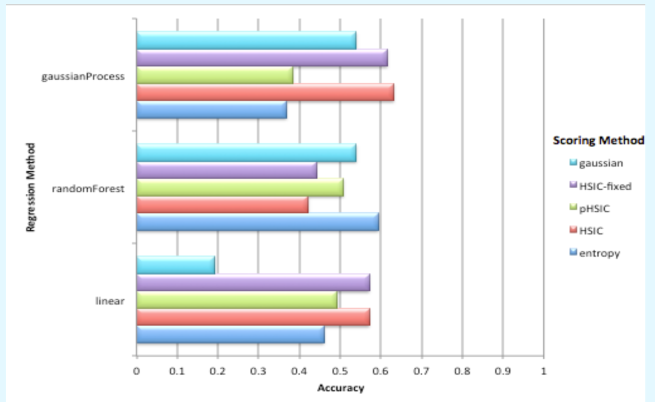


We consider two different transformation selection schemes. The first minimizes the mean squared error of the regression model, and the second applies Tukey's transformation to both *X* and *Y* to reduce the skew for each variable.

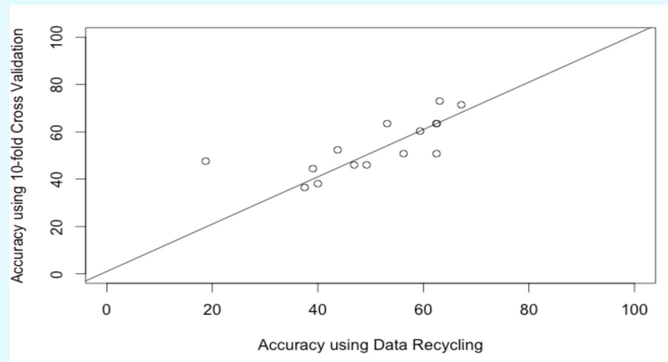


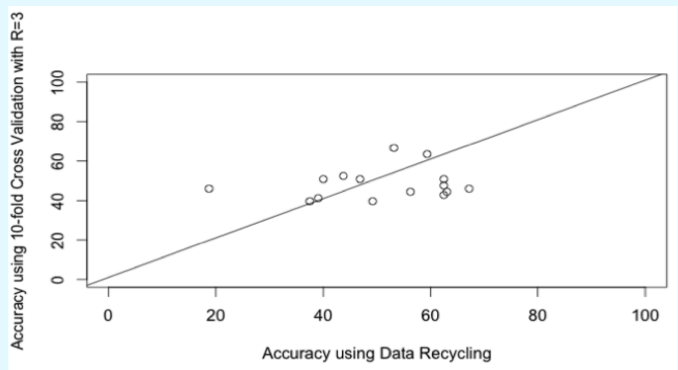
5. Experiments

We compare performance of bivariate ANMs using gaussian process regression, random forest regression and least squares regression as a baseline.

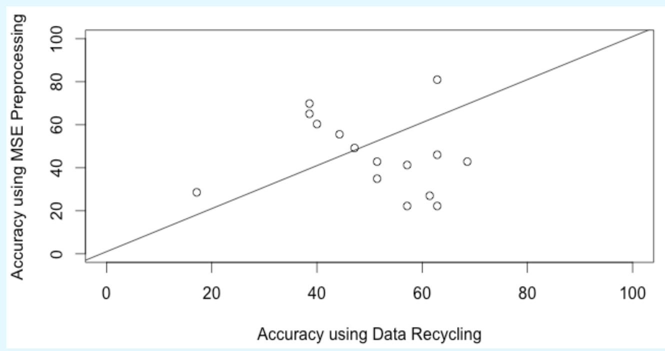


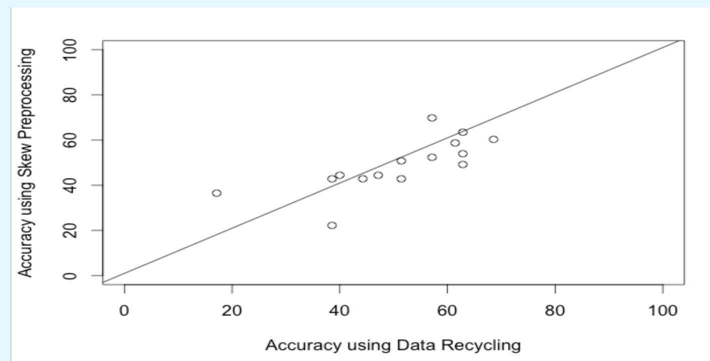
We compare performance of bivariate ANMs using **data recycling**, or the same data for both training and testing, and using 10-fold cross-validation.





We used Tukey's Ladder of Powers to preprocess data before regression





6. Real data and contributions

Experiments using real data

We applied variations of bivariate ANM to the Cause-Effect Pairs benchmark dataset [2]. This is the same dataset used to obtain results in [1]. We used a subset of the CEP dataset including pairs with less than 2000 data points.

Most of these pairs are taken from sources where we can be fairly confident in the given ground truth, e.g. weather measurements. However, the dataset violates assumptions required for bivariate ANM. Some pairs have known confounding factors, or are taken from time series autocorrelated variables with feedback relationships. Thus, dataset applicability is a threat to validity affecting all results shown.

Contributions

We have preliminary results indicating the behavior of the bivariate causal inference algorithm using ANMs has highly variable accuracy. The causal direction inferred with ANMs is highly sensitive to the regression and scoring methods used, whether residuals are obtained using data recycling, and the method of data preprocessing.

[1] Mooij, J. M., Peters, J., Janzing, D., and Zscheischler, J. (2014). Distinguishing cause from effect using observational data: methods and benchmarks. arXiv preprint arXiv:1412.3773.
[2] Zscheischler, J. (2015). Benchmark data set for causal discovery algorithms. Retrieved from http://webdav.tuebingen.mpg.de/cause-effect/
[3] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[4] Scott, D. (2013). "Transformations: Tukey's Ladder of Powers." Retrieved from http://onlinestatbooks.com/2/transformations/tukey.html, May 2015.