Towards Evaluating and Understanding the Additive Noise Model for Social Science

Kwonsoo Chae, Peerapon Akkapusit, Jaeryoung Ka

Abstract

Causality analysis is a key tool for understanding and explaining the actual workings of multiple factors in diverse scientific applications. It is useful not only because it enables us to pinpoint the precise causes of undesirable phenomena, but also because it permits policy makers to substantiate their claims and decisions with scientific evidence. Existing machine learning techniques, however, often fail to consider causal relationships because their foundation is based on the concept of association, not causation. In this project, we aim at studying the techniques (mostly ANM) for identifying right causal directions given observational data. We pose three research questions, and answer them with experimental justification.

1 Introduction

Understanding causal relationships among distinct factors is one of the ultimate goals of scientific studies such as econometrics, political science, psychology, sociology, and education. Comprehension of underlying causal mechanisms permits researchers and practitioners to enjoy at least two benefits. The first is that it enables us to target proper factors so that the problem of interest may actually be solved. Let us suppose that there is a positive correlation between ice cream sales and the number of people who drown in the summer. Reducing the ice cream sales do not reduces the number of people who drown, since there is an underlying causality factor "summer" that actually causes both factors mentioned earlier. In this case, if we do not know this underlying mechanism, we may decide to focus on a wrong factor (i.e., reducing the ice cream sales) that may not fix the issue of interest. The second is that it enables us to substantiate our claim properly. Specifically, when people try to solve an important issue of their own society, they often need to provide proper evidence for their decisions. For example, suppose that we are elementary school teachers and we believe that playing computer games more than five hours a week has negative effects on their studies. We want to make a new school policy that prevents students from spending more than five hours a week playing computer games. To persuade the students and their parents, it is important to substantiate our claim using scientific evidence that spending more than five hours actually has negative effects. In this case, understanding the causal relationship between the hours on computer games and students' performances will greatly help.

Most of the existing machine learning techniques, however, are based on correlational association rather than causal relationships. The main interest of the techniques is to find useful patterns of the world from the given data, not to find actual causal relationships among distinct factors. Let us compare supervised learning and causality analysis. The goal of supervised learning is to compute $p(X|Y = y)$ where X and Y are random variables and y denotes the actual data that we observed. Informally, computation of the quantity lets us answer what we can know about the factor X when we already know some information (i.e., y) about Y, and this is based on the concept of association. In causality analysis, on the other hand, our goal is to compute the effect of the intervention and we can formulate the quantity as $p(X|do(Y = y))$ using the *do* operation. Informally, it enables us to answer how artificially setting Y to y affects the factor X, and this is based on causality, not an association. The causality theory discharges such computations and provides guarantees and limitations.

In this project, we study two techniques, the additive noise model (ANM) [1] and information geometric causal inference (IGCI) [2] (paying more attention to ANM), for identifying causal directions. In the development of the causal theory [3], the theory provides proof that the directional information cannot be identified only using observational data; the knowledge of causal directions (such as "X causes Y") is essential for the theory. Then, the theory itself places great emphasis on computing the causal effect (i.e., $p(X|do(Y = y))$) assuming that the directional information (i.e., "Y causes X") is given. In practice, however, discovering the causal direction is

nontrivial and important to solve the problem in diverse social science applications. We focus on finding out directional information, and formulate three research questions as follows:

- RQ1 (Replication): Can we reproduce the ANM results reported by [4]? Is there any issue?
- RQ2 (Possible reasons for incorrect prediction): Can we identify why ANM fails to produce correct causal directions on some datasets? Can we provide a good explanation?
- RQ3 (Extended, categorized evaluation): How do ANM and some other model (IGCI) perform when we categorize datasets (including new ones) according to several different social science fields?

The research questions are the key ingredients of our contribution, and we answer the questions in Section 5 including the experimental results.

2 Related Work

Researchers have developed diverse techniques for identifying causal directions given observational data. One way of classifying the methods is to check if a method is a pairwise method (i.e., basically for bivariate datasets as in our project) or if a method is for identifying the graph G at ones. CGNN [5], GES, GIES, LiNGAM, and CAM [6] are for discovering a directed acyclic graph G, which represents the whole causality directions among the variables (represented as nodes in the graph). ANM [1], IGCI [2], RCC [7], NCC [8], GNN [5], GPI [9], PNL [10], and Jarfo [11] are the methodologies that mainly identifying causal directions for bivariate cases. Among the techniques, our project focuses on ANM (and IGCI), and studies its models and ideas.

Mooij et al. [4] carried out in-depth evaluation of ANM and IGCI, two popular methods for identifying causal directions, and more importantly, provided a useful benchmark for evaluating methodologies for identifying directional information in the bivariate cases. Their work motivated our project. Our project reproduces it results on ANM, and we take one step further. Specifically, we extended the datasets by collecting more datasets from Kaggle [12] focusing only on social science applications, categorized them according to several important subfields of social science, and evaluated ANM in this extended, categorized setting.

3 Data

We utilized 300 bivariate data sets in this paper. First, we borrowed 108 bivariate data sets from our target paper, [4]. We utilized **96** out of 108 data sets from the paper, excluding 52, 53, 54, 55, 71, 81, 82, 83, 105 for the reason of being non-bivariate and 58, 86, 104 for unspecified ground truth. Second, we imported 4 Kaggle data sets [13] [14] [15] [16] and pick columns out of the data sets to create a bivariate data sets with specific ground-truth which is, in total, 214 data sets.

3.1 Preprocessing data

Since the 96 bivariate data sets from previous paper [4] are not meaningful for analyzing ANM towards each category of social science data set, we utilized 210 data sets from kaggle which are categorized to 4 social science data including

1. Category 1: Health and public welfare - 56 data sets
   Column name and descriptions:
   - Happiness score: A metric measured in 2015 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
   - Economy: The extent to which GDP contributes to the calculation of the Happiness Score.
   - Family: The extent to which Family contributes to the calculation of the Happiness Score
   - Health: The extent to which Life expectancy contributed to the calculation of the Happiness Score
   - Freedom: The extent to which Freedom contributed to the calculation of the Happiness Score.

- Trust (Government Corruption): The extent to which Perception of Corruption contributes to Happiness Score.
- Generosity: The extent to which Generosity contributed to the calculation of the Happiness Score.
- Dystopia Residual: The extent to which Dystopia Residual contributed to the calculation of the Happiness Score.

2. Category 2: Politics and policies - 50 data sets
   Column name and descriptions:
   - imonth: the number of the month in which the incident occurred.
   - iday: the numeric day of the month on which the incident occurred.
   - latitude: the latitude of the city in which the event occurred.
   - longitude: the longitude of the city in which the event occurred.
   - region: the field identifies the region code based on 12 regions (0 to 11)

   We further divided data sets by decades in which the incident occurred, yielding 5 data sets per each above. (1970s to 2010s)
3. Category 3: Humanities - 52 data sets
   Column name and descriptions:
   - yr_ built: Built Year
   - grade: overall grade given to the housing unit, based on King County grading system
   - sqft_above: square footage of house apart from basement
   - floors: total floors (levels) in house
   - sqft_living: square footage of the home
   - bedrooms: number of Bedrooms/House
   - bathrooms: number of bathrooms/House
   - price: price of the house

   We further divided data sets by the time of transaction, yielding two (2014.3 - 8, 2014.9 - 2015.2) data sets per each above.
4. Category 4: Economics - 56 data sets
   Column name and descriptions:
   Set 1:
   - pf_rol: Index of Rules and Law
   - pf_ss: Index of Security and safety
   - pf_religion: Religious freedom
   - pf_assiociation: Freedom to associate and assemble with peaceful individuals or organizations
   - pf_expression: Freedom of expression
   - pf_score: Personal Freedom (score)

   Set 2:
   - ef_government: Size of government
   - ef_legal: Index of Legal system and property rights
   - ef_money: Index of sound money
   - ef_trade: Freedom to trade internationally

- ef_regulation: Index of Regulation
- ef_score: Economic Freedom (score)

Set 3:

- ef_trade_movement: Controls of the movement of capital and people
- ef_trade_tariffs: Index of tariffs
- ef_trade_black: Black market exchange rates
- ef_trade_regulatory: Index of Regulatory trade barriers
- ef_trade: Freedom to trade internationally

Set 4:

- ef_regulation_credit: Credit market regulations
- ef_regulation_labor: Labor market regulations
- ef_regualtion_business: Business regulations
- ef_regulation: Index of Regulation

Set 5:

- pf_score: Personal Freedom (score)
- ef_score: Economic Freedom (score)
- hf_score: Human Freedom (score)
- hf_quartile: Human Freedom (quartile)

We divided columns further by choosing all possible bivariate data sets from each pair of column.

4 Methods and Techniques

In this section, we review the additive noise model (ANM) introduced by Hoyer et al. [1], and information geometric causal inference (IGCI) introduced by [2]. After the presentation of ANM the two techniques, we also explain our simple method for automatically generating ground truth for additional datasets that are used in our extended evaluation (in Section 5.3).

4.1 Additive noise model (ANM)

We review ANM presented in [1]. The authors proved (In section 3 of the paper) that one of "x causes y" and "y causes x" may be determined under some conditions (e.g., that the true model is identifiable). They also demonstrated (in Section 4 of the paper) a practical method for actually identifying the true model. Our focus is to understand the modeling assumptions and the identification procedure.

Modeling assumptions Suppose that there is a data generation process, and we represent it by a directed acyclic graph G. Let the random variables $x_1, \ldots, x_n$ are observed. In the additive noise model, the observed variables are assumed to have the following relationships:

$$x_i := f_i(x_{\mathrm{pa}(i)}) + n_i$$

where $x_{\mathrm{pa}(i)}$ are the parents of x_i in the graph G, and each n_i is an independent noise, that is, $p_{\mathbf{n}}(\mathbf{n}) = \prod_{i=1}^{n} p_{n_i}(n_i)$, where $\mathbf{n}$ is the vector of $x_1, \ldots, x_n$. The functions f_i and the distributions $p_{n_i}(n_i)$ can be arbitrary. We observe the data $\mathbf{x} = (x_1, \ldots, x_n)$ possibly multiple times. The modeling considers that the data are generated (sampled) independently using G together with the same distributions specified in the model. Then, the main goal is to infer the graph G given the data.

A special case of the model is the nonlinear Gaussian case. Suppose there are two observed variables x and y. Then, the model is as follows:

$$y := f(x) + n$$

where x and n are independent and both Gaussian. Though x and n are both Gaussian, y is not since the function f is nonlinear. In this case, inferring the graph G reduces to differentiating between "x causes y" and "y causes x".

Identification procedure This paragraph describes how ANM determines the true causal graph G. Consider the bivariate case with x and y denoting the two variables.

Algorithm 1 Identification procedure in [1].

Input: two observable variables x and y, two causal models $y := f(x) + n$ and $x := g(y) + n$ with arbitrary functions f and g.

Output:

1: **if** is_independent(x, y) **then return** "No causal relationship." ▷ Independent test
2: **else**
3: $\hat{f} \leftarrow$ nonlinear_regression(f, y, x) ▷ Nonlinear regression of y on x to get an estimate $\hat{f}$ of f
4: $\hat{n_1} \leftarrow y - \hat{f}(x)$ ▷ Compute residuals
5: $i_1 \leftarrow$ is_independent$(\hat{n_1}, x)$ ▷ Independence test
6: $\hat{g} \leftarrow$ nonlinear_regression(g, x, y) ▷ Nonlinear regression of x on y to get an estimate $\hat{g}$ of g
7: $\hat{n_2} \leftarrow x - \hat{g}(y)$ ▷ Compute residuals
8: $i_2 \leftarrow$ is_independent$(\hat{n_2}, y)$ ▷ Independence test
9: **if** i_1 and i_2 **then** "Either may be correct. Cannot infer from the data."
10: **else if** i_1 **then** "x causes y."
11: **else if** i_2 **then** "y causes x."
12: **else** "Generating mechanism more complex. Cannot be described by the model."

Algorithm 1 describes the identification procedure. In the procedure, line 3 through 8 are the core steps and the final decision of ANM (line 9 through 12) is directly depends on the results of the core steps. Line 3 through 5 are for the first causal model $y := f(x) + n$, and line 6 through 8 for the other causal model $x := g(y) + n$. The variables i_1 (line 5) and and i_2 (line 8) store the results for the first and second directions, respectively. Let us focus on the first causal model (i.e., $y := f(x) + n$) to describe what the key steps are (line $3 - 5$). It first performs nonlinear regression of y on x to compute an estimate $\hat{f}$ of f (line 3). Then, it calculates the corresponding residuals (estimation errors) $\hat{n_1}$ (line 4). Finally, it performs an independence test (a kind of hypothesis testing) to check if the residuals are independent of x. Then, it stores the binary results (i.e., whether the two quantities are independent or not)in the variable i_1. It repeats the same steps for the second causal model $x := g(y) + n$. The primary purpose of the key steps (line $3 - 5$ for the first causal model and line $6 - 8$ for the second) is to test whether the model is consistent with the given data observations. Hoyer et al. [1] used Gaussian Processes [17] for nonlinear regression, and kernel methods [18] for independence tests. They reported that the kind of techniques for the regression and the independence test did not significantly affect the overall results.

4.2 Information geometric causal inference (IGCI)

In this part, we will take a look at Information Geometric Causal Inference with the noiseless model, or IGCI, presented by 2 papers [2], [19]. The main idea is that one of "x causes y" and "y causes x" can be determined by investigating the slope of function whether it correlates to the peak of the density of each variable or not. In section 2 of the paper [19], they revisited additive noise models and showed that entropies can be a key role for describing dependencies between densities. Then, they developed the information geometric perspective in section 3 and results in section 4 [19]. Again, our main goals are to comprehend their procedure described in Section 4 of the paper [19] and to apply our social science data based on the key quantities in the procedure.

The main idea of IGCI is as followed

1. Problem: we infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the correct causal model

2. Main Focus: If $X \to Y$ then f and P_x (marginal distribution of X) are independent, i.e, peaks of P_x do not correlate with slope of f
3. Hence: peaks of P_y correlate with slope of f^{-1}

Let $Y = f(X)$, where f is a monotonously increasing differentiable bijection of [0,1] with differentiable inverse f^{-1} If $logf^{'}$ and P_x are uncorrelated, in other words,

$$\int logf^{'}(x)P(X)dx = \int logf^{'}(x)dx$$

then $log(f^{-1})^{'}$ and P_Y are positively correlated, i.e.,

$$\int log(f^{-1})^{'}(x)P(y)dy > \int logf^{'}(y)dy$$

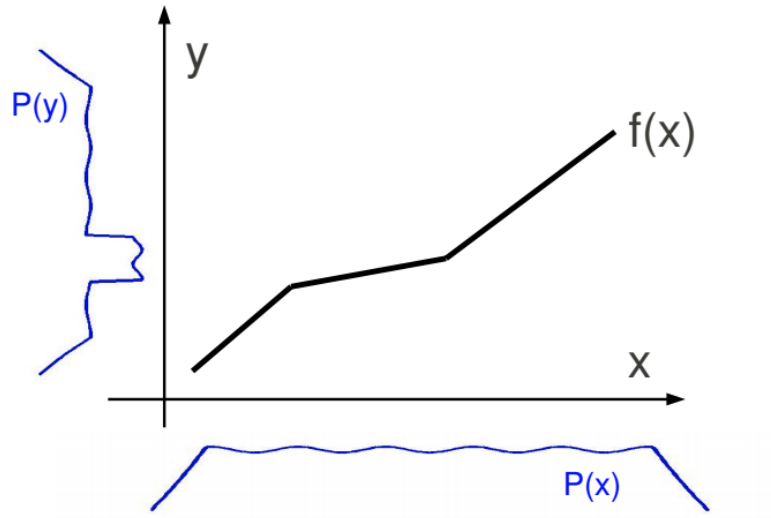

Figure 1: If the structure of the density of P_X is not correlated with the slope of f, then flat regions of f induce peaks of P_Y. The causal hypothesis $Y \to X$ is thus impossible because the causal mechanism f^{-1} appears to be adjusted to "input" distribution P_Y, i.e the change of value Y does not affect the change of value X

Slope-based IGCI infer $X \to Y$ whenever

$$\int_0^1 log|f^{'}(x)|P(x)dx < \int_0^1 log|f^{-1}|^{'}(y)P(y)dx$$

(High density P_Y tends to occur at large slope of f^{-1}) And the estimator is then introduced as followed:

$$\hat{C}_{X\to Y} := 1/m \sum_{j=1}^{m} log|(y_(j+1) - y_j)/(x_(j+1) - x_j)| \approx \int log|f^{'}(x)|P(x)dx$$

We infer $X \to Y$ whenever

$$\hat{C}_{X\to Y} < \hat{C}_{Y\to X}.$$

which is our main focus on information geometric causal inference (IGCI). With this information we have, we can conclude causal direction using this estimator ($C_{X\to Y}$ or $C_{Y\to X}$) we calculated with integral approximation estimator.

4.3 Generation of ground truth

In Section 5.3, we perform extended experiments with categorized social science datasets. For the experiments, We collected additional datasets from Kaggle [12]. We were not able to avoid the additional data collection since we thought that the number of datasets in each category was not enough for a valid evaluation. New datasets from Kaggle helped us solving this data size issue, but introduced another challenge: the new datasets are not equipped with causal ground truth. For evaluation, having ground truth is essential, and so we had to come up with a technique for automatically labeling the causal ground truth in each dataset.

Our idea is to use multiple methodologies for identifying causal directions and consider the majority of the decisions as ground truth. The available techniques were as follows:

1. ANM (Additive Noise Model) [1]
2. IGCI (Information Geometric Causal Inference) [2]
3. Bivariate fit model
4. CDS (Conditional Distribution Similarity Statistic) [11]
5. GNN (Generative Neural Networks for causal inference (pairwise)) [5]
6. RECI (Regression Error based Causal Inference) [20]

Among the six techniques, we excluded ANM and used the other five methods to generate ground truth. Excluding ANM is a reasonable choice, since what we wanted to do is the evaluation of ANM and we wanted the ground truth not to depend on the ANM technique itself. Likewise, we also did the same to IGCi by excluding IGCI when we created ground truth for IGCI and included ANM.

5 Results

In this section, we answer our research questions together with experimental results. For the readers, we repeat our research questions below:

- RQ1 (Replication): Can we reproduce the ANM results reported by [4]? Is there any issue?
- RQ2 (Possible reasons for incorrect prediction): Can we identify why ANM fails to produce correct causal directions on some datasets? Can we provide a good explanation?
- RQ3 (Extended, categorized evaluation): How do ANM and some other model (IGCI) perform when we categorize datasets (including new ones) according to several different social science fields?

We answer the research questions in the following subsections.

5.1 Results for RQ1

We tried to reproduce the performance of ANM in terms of accuracy reported in [4]. For this, we performed an experiment using an ANM implementation [21] and the benchmark datasets provided by Mooij et al. [4] (and organized in a website [22]). The benchmark has in total 108 datasets, but datasets 52, 53, 54, 55, 71, 81, 82, 83, 105 are nonbivariate, and datasets 58, 86, 104 are not equipped with ground truth. For this reason, we only used 96 dataset out of the 108 datasets. Our evaluation results showed that the overall accuracy on the 96 datasets was 53.16%. The reported accuracy from the paper by Mooij et al. [4] was $63 + -10\%$, and so we checked that we were able to reproduce the results. We additionally calcuated recall and precision on the whole $96 datasets$, since they are the most popular evaluation criteria in machine learning generally. The precision was 72.2% and the recall was 56.5%.

While working with the ANM implementation, we found that the decision process of the implementation is somewhat different from what the original ANM paper [1] says. Specifically, in the ANM identification procedure, there are four different possibilities in the bivariate case with the variable X and Y: "X causes Y", "Y causes X", "both of the directions", and "neither of the directions". The last two cases may happen either because the modeling

assumptions may not be compatible with the observed data, or because we need more compex models. In the actual implementation [21], however, there may be only two cases: "X causes Y" and "Y causes X". They compared the independence scores of the two causal directions, and make the final decision based on the sign of the score difference. This is not in line with the original ANM procedure, because the original one makes decisions on each direction separately. We also tried to fix the implementation so that it becomes in line with the original ANM methodology, and re-evaluate it on our dataaset, but we eventually failed to do so.

5.2 Result for RQ2

After the replication experiments, we take one step further to investigate why ANM did not make correct predictions on some datasets. This question is important for coming up with an improvement on ANM in the future. Here we remind the readers of the fact that the ANM procedure (for both directions) first solve a nonlinear regression problem, and then test if the residuals and the noise are really independent (as the model assumed). We designed a simple experiment, where we not only checked the final ANM decision results but also checked the intermediate performances of the nonlinear regression. Our initial conjecture was that the poor performance in the regression problem may affect the final ANM performance seriously. For the performance of the intermediate regression, we checked the size of residuals (i.e,. $\hat{f} - f$). From the simple experiments, we found that the intermediate regression performance may actually have strong effects on the final ANM performances.

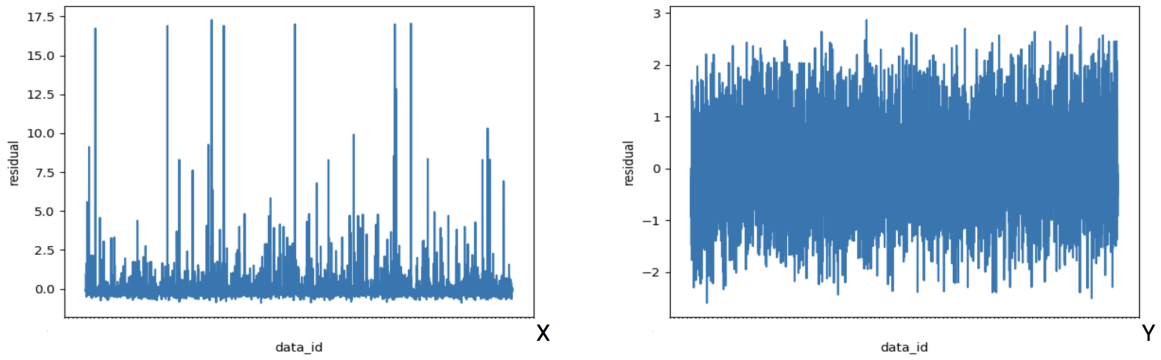

Figure 2: Residual (y axis) and data id (x axis) representing true direction from X to Y (left) and ANM wrong prediction from Y to X (right)

Figure 2 shows the residuals from the regression in both causal directions for the dataset 17. The graph on the left shows the regression results of the true causal direction ($X \to Y$), and the graph on the right plots the regression results of the wrong causal direction ($Y \to X$), which our ANM procedure incorrectly chose. The major difference shown in the two graphs is that the regression performance of the true direction (i.e., the graph on the left) is more poor than that of the wrong direction (i.e., the graph on the right); we can observe easily observe the residuals with large absolute values. On the other hand, the residuals shown in the second plot have small values. This may be problematic, since the ANM procedure checks if the residuals and the noise are independent to make the final decision. The large residual values may introduce dependencies between the residuals and the noise, which may lead us to wrong predictions.

For an easy check, we tried artificially setting all the residuals values to 0 in the true direction for the datasets that are not correctly predicted. This is because we wanted to see if the ANM predicts the direction correctly when we perfectly solved the intermediate regression problem and so it does not introduce any unnecessary dependencies between residuals and noises. Indeed, we ended up seeing that all the dastaset were predicted correctly in this artificial setting. This gives us a lesson that when we apply the ANM methodology to any social science applications, we need to be careful about the intermediate regression performance before using the ANM results. [4] mentions the similar point in their paper.

5.3 Result for Research Question 3

To answer RQ3, we divided our datasets into four different categories: health and public welfare, politics and policies, humanities, and economics. We made this categorization, since they are major subfields of social science, and it is useful to know how ANM performs in the four different subfields. An issue here was that after categorization the numbe of dataset in each category became much smaller, and so the validness of evaluation results became questionable. To resolve this issue, we additionally collected more datasets outside of the paper by [4]. Specifically, we found more datasets from Kaggle [13, 14, 15, 16], and made the numbe of dataset in each category larger. In total, we used 214 social science datasets from Kaggle, and 96 datasets that we originally have. Out of the original 96 datasets, 19 datasets are about social science, and 77 datasets are about non-social science. Our study about RQ3 therefore is not just about categorized evaluation with extended datasets, but also about focusing more on sicial science datasets.

Table 1: Performance of ANM for different social science data type

Category 1: Health and Public Welfare	Precision	$X \to Y$:75.0% (out of 56)	$Y \to X$:77.8% (out of 56)
	Recall	$X \to Y$: 75.0% (out of 56)	$Y \to X$:77.8% (out of 56)
	Accuracy	76.5% (out of 56)	
Category 2: Politics and Policies	Precision	$X \to Y$:20.0% (out of 50)	$Y \to X$:47.8% (out of 50)
	Recall	$X \to Y$:33.3% (out of 50)	$Y \to X$:23.6% (out of 50)
	Accuracy	47.1% (out of 50)	
Category 3: Humanities	Precision	$X \to Y$: 63.6% (out of 56)	$Y \to X$:66.7% (out of 56)
	Recall	$X \to Y$: 73.6% (out of 56)	$Y \to X$:55.6% (out of 56)
	Accuracy	70.6% (out of 56)	
Category 4: Economics	Precision	$X \to Y$: 44.4% (out of 52)	$Y \to X$:36.4% (out of 52)
	Recall	$X \to Y$: 36.4% (out of 52)	$Y \to X$:44.4% (out of 52)
	Accuracy	47.1% (out of 52)	

Table 2: Performance of IGCI for different social science data type

Category 1: Health and Public Welfare	Precision	$X \to Y$: 61.9% (out of 56)	$Y \to X$:81.3% (out of 56)
	Recall	$X \to Y$: 81.3% (out of 56)	$Y \to X$:61.9% (out of 56)
	Accuracy	70.3% (out of 56)	
Category 2: Politics and Policies	Precision	$X \to Y$:28.6% (out of 50)	$Y \to X$:46.7% (out of 50)
	Recall	$X \to Y$: 36.4% (out of 50)	$Y \to X$:58.3% (out of 50)
	Accuracy	48.6% (out of 50)	
Category 3: Humanities	Precision	$X \to Y$: 39.1% (out of 56)	$Y \to X$:35.7% (out of 56)
	Recall	$X \to Y$: 50.0% (out of 56)	$Y \to X$:26.3% (out of 56)
	Accuracy	37.8% (out of 56)	
Category 4: Economics	Precision	$X \to Y$: 42.1% (out of 52)	$Y \to X$:33.3% (out of 52)
	Recall	$X \to Y$: 44.4% (out of 52)	$Y \to X$:31.3% (out of 52)
	Accuracy	35.1% (out of 52)	

We applied ANM and IGCI to the extended, categorized datasets. Table 1 and Table 2 show the results. In terms of accuracy, ANM demonstrated good performance on the first and third categories (accuracy 76.5% and 70.6%, respectively), and IGCI performed well only for the first category. The ANM and IGCI are both applied to each type of social science dataset processed from section 3. As the tables show, both ANM and IGCI demonstrated good performance for Health and Public Welfare data (Category 1). However, only ANM performed well on Humanities data (Category 3), and both performed bad for Politics and Policies data(Category 2) and Economics(Category 4) data. We can see that ANM demonstrated different performances on difference social science datasets, and different methodologies (ANM and IGCI) produce difference performances depending on the category. A lesson we take from here is that investigating the data patterns before application of a specific methodology for finding out the right causal direction may help make correct predictions.

6 Discussion and conclusion

The discovery of directional information does not receive much attention in the development of causality theory [3]; the theory assumes that the directional information is given, and focuses on the computation of causal effects. In the bivariate case, this means that the theory itself focuses on computing $p(X|do(Y = y^*))$[1], the distributional change on the random variable X when we intervene and forcefully fix the random variable Y to a specific value y^*, assuming that either $X \to Y$ or $Y \to X$ are given. In practical settings, however, identifying the directional information is very important to actually solve a problem relevant to causality. In fact, in diverse applications, figuring out such directional facts is not trivial even on a simple dataset.

Researchers have already developed various techniques [1, 2, 11, 5, 20, 6, 9, 10] for identifying causal directions. In this project, we studied two techniques (ANM and IGCI, but paying more attention to ANM) for the identification of causal directions focusing on social science data. Although we tried to analyze a possible reason for incorrect decisions of ANM and to evaluate ANM (and IGCI) in a more systematic setting focusing on social science applications, we still do not fully understand the theoretical implications of ANM and other existing techniques. Studying their theoretical properties, different modeling assumptions, and more importantly, their own limitations may help us find some possibilities of coming up with a new idea for improving the performance of the methodologies.

Causality analysis plays a key role in diverse social science applications, since it enables researchers and practitioners to pinpoint the underlying causal mechanisms and effects among different factors. Understanding causal mechanisms is particularly important both when people need to find the real causes of a social phenomenon and when people want to substantiate their claims. In this project, we studied and evaluated ANM (plus IGCI), a technique for identifying causal directions given observational data. We believe that it is a good start in further studying causality techniques and applying them to social science applications.

References

[1] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[2] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[3] Judea Pearl. *Causality*. Cambridge university press, 2009.

[4] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[5] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

[6] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[7] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

[8] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017.

[1] We do not discuss more details on the *do* operation in this report. It is enough to know that it means intervention in the data generation process.

[9] Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in neural information processing systems*, pages 1687–1695, 2010.

[10] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

[11] José AR Fonollosa. Conditional distribution variability measures for causality detection. *arXiv preprint arXiv:1601.06680*, 2016.

[12] Kaggle, . URL https://www.kaggle.com.

[13] Global terrorism database, . URL https://www.kaggle.com/START-UMD/gtd.

[14] World happiness report, . URL https://www.kaggle.com/unsdsn/world-happiness.

[15] House sales in king country, . URL https://www.kaggle.com/harlfoxem/housesalesprediction.

[16] The human freedom index, . URL https://www.kaggle.com/gsutters/the-human-freedom-index.

[17] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.

[18] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.

[19] Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

[20] Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schoelkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909, 2018.

[21] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

[22] Bivariate causal direction benchmark. URL http://webdav.tuebingen.mpg.de/cause-effect/.