A Novel Principle for Causal Inference in Data with Small Error Variance

Patrick Blöbaum, Shohei Shimizu, Takashi Washio

The Institute of Scientific and Industrial Research, Osaka University

Abstract. Causal inference addresses the problem of identifying cause and effect variables in observed data. While most of the current techniques base heavily on exploiting asymmetries in the error noise, these techniques struggle in data that only contain small noise. We present a novel principle for causal inference in a bivariate setting with small error variance. For this, we exploit an asymmetry in the prediction error under the assumption of additive noise and an independence between data generating mechanism and its input. The applicability of our approach is corroborated with empirical evaluations in artificial and real-world data sets.

1 Introduction

Causal inference is an increasingly popular topic that is about determining the causal relationship between measured variables. For example, in observing the development of a disease, we could ask whether a certain symptom causes the disease or the disease causes the symptom.

This paper addresses the setting of causal inference in data that only have a small variance in the error noise. Data with small error noise are particularly difficult in the domain of causal inference, since many common techniques are based on exploiting the noise [1, 2] and are struggling if the noise is too small. Our work is based on [3], who formulate the theorem that the expected prediction error heavily depends on the causal direction between two variables, i.e. predicting the effect from its cause leads to a smaller error than predicting the cause from its effect. We show that this asymmetry holds for the leastsquares solutions in both possible causal directions when the noise is small, and thus, provide a simple applicable causal inference algorithm that exploit this asymmetry. The theorem has two main assumptions; an additive noise affecting the effect and an independence between cause distribution and data generating function. These assumptions are further explained in the next section.

2 Problem Setting, Assumptions and Theory

We consider a two variable model with an observable *cause* variable C, an observable *effect* E and a latent noise variable N_E with $\mathbb{E}[N_E] = 0$. The noise is assumed to be independent of the cause $C \perp N_E$ and all densities are assumed to be strictly positive.

Causal Inference Given data X and Y sampled from the joint distribution p(X, Y), the problem setting of causal inference is to infer which variable is the cause C and which variable is the effect E.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Additive Noise Model For the theoretical background, we consider the additive noise model (ANM) [1] that assumes the effect data were generated with an additive noise term $E = \phi(C) + N_E$, where $C \in \mathbb{R}$, $E \in \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ is a non-linear diffeomorphism. ϕ is assumed to be normalized without loss of generality since it is only a scaling problem.

Independence assumption Another crucial assumption is the *independence of mechanism and cause*. This assumption states that the shape of ϕ is designed independently of the cause distribution p(C). More details about a formalization and corresponding implications can be found in [4]. This assumption can also be seen in the way that changing the distribution p(C) does not change the function ϕ and vice versa. This is reasonable, e.g. considering that the most physical processes can be precisely described in a mathematical form that is universally valid in any environment.

Error Asymmetry As already mentioned before, the fundamental theoretical background of our algorithm is based on the error asymmetry theorem by [3]. Under the ANM and independence assumption, this theorem states that the expected squared error of predicting the cause from the effect is generally smaller than or equal to the error of predicting the effect from the cause with respect to the data generating function $\phi: \mathbb{E}[(\phi(C) - E)^2] \leq \mathbb{E}[(\phi^{-1}(E) - C)^2]$.

3 Related Work

Widely used techniques for causal inference are based on the ANM, where in causal direction, the prediction error is assumed to be independent of the input [1]. Another approach is LiNGAM [2] which assumes that ϕ is linear and N_E non-Gaussian, which then allows the identification of the causal direction via independent component analysis. A more recent approach by [5] addresses the deterministic case $E = \phi(C)$ without noise and uses the independence assumption to infer the causal direction by checking which direction fulfills the assumption.

Our approach exploits the aforementioned error asymmetry which combines the ANM and independence assumption. We address the case of small error noise where our approach does not rely on any independence tests. Furthermore, fitting a model in the data and comparing the squared errors, depending on the regression model, have a generally low computational cost.

4 Algorithm

Following the theoretical background given in Section 2, the algorithm is straightforward. Given data X, Y sampled from p(X, Y), the key idea is to fit a regression model in both possible causal directions, i.e. $f: X \to Y$ and $g: Y \to X$. As stated in [3], due to the additive noise, the general problem of estimating ϕ and its inverse ϕ^{-1} is that the ordinary least-squares solution for ϕ^{-1} can be highly biased, but is generally unbiased for ϕ . Seeing that $\arg\min_{f} \mathbb{E}[(f(C) - \phi(C) -$ ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

 $N_E)^2] = \phi$ and $\lim_{\operatorname{Var}[N_E]\to 0} \underset{g}{\operatorname{arg\,min}} \mathbb{E}[(g(\phi(C) + N_E) - C)^2] = \phi^{-1}$, we justify the usage of the least-squares solution of f and g by taking a look at the error asymmetry for the limit of $\operatorname{Var}[N_E] \to 0$. It can be concluded that

$$\lim_{\operatorname{Var}[N_E]\to 0} \frac{\mathbb{E}[(g(\phi(C)+N_E)-C)^2]}{\mathbb{E}[(f(C)-\phi(C)-N_E)^2]} = \lim_{\operatorname{Var}[N_E]\to 0} \frac{\operatorname{Var}[N_E] \int \left(\frac{1}{\phi'(C)}\right)^2 p(C) dC}{\operatorname{Var}[N_E]}$$
$$= \int \left(\frac{1}{\phi'(C)}\right)^2 p(C) dC \ge 1 \tag{1}$$

where we used the property that $f = \phi$ and $g = \phi^{-1}$ for $\operatorname{Var}[N_E] \to 0$ and we know from the error asymmetry theorem that $\int (\frac{1}{\phi'(C)})^2 p(C) dC \geq 1$ with equality only if ϕ is linear [3]. Therefore, if ϕ is non-linear, the causal direction can then be determined by comparing the expected squared errors of f and gaccording to (1). From the theoretical perspective, it can be expected that our approach performs the best in data with a low noise variance. Although a statement for data with a high noise variance is not possible, we observed that the algorithm performed well even in these data. Our proposed algorithm is summarized in Algorithm 1 where the regression models are assumed to be capable of approximating ϕ . Note that the assumption about ϕ to be a diffeomorphism is convenient for assuring the existence of ϕ^{-1} , but if the inverse does not exist, we can particularly expect a higher error for the inverse direction.

5 Experiments

We evaluated our algorithm on artificial and real-world data sets. The performance were compared with three different commonly used causal inference techniques:

ANM: The classical ANM [1] approach tests for independence between input and prediction error. In our evaluations, we tested independence via the Hilbert-Schmidt independence criterion.

IGCI: The information geometric causal inference (IGCI) [5] approach is based on the independence assumption and checks in which causal direction is the input distribution uncorrelated with the functional shape.

LiNGAM: The Linear Non-Gaussian Acyclic Model (LiNGAM) [2] assumes a linear relationship with non-gaussian noise, which allows an identification of the causal direction.

In case of the ANM and our approach, we utilized a smoothing spline as regression model, because it is less sensitive to the choice of parameters than e.g. Gaussian Process Regression. Generally, each evaluation was averaged over 100 runs. Note that we forced a decision from all algorithms, even if e.g. the ANM causal inference algorithm would reject a decision regarding an independence. In these cases, the more likely causal direction based on the test statistics was taken.

The additive noise assumption is in common with the ANM approach and the independence assumption is in common with the IGCI approach. While the main drawback of ANM and LiNGAM is that the noise in the data is too weak, a general drawback of IGCI is that the underlying function needs to be sufficiently non-linear and this also needs to be captured in the observed data. However, combining the additive noise and independence assumption in our method provides several advantages to overcome these problems.

Since we address data with small error noise, an answer to the question of how much smaller $\operatorname{Var}[N_E]$ should be compared to $\operatorname{Var}[\phi(C)]$ is difficult and heavily depends on the underlying problem. Therefore, in our experiments, we set the constrain

$$\operatorname{Var}[N_E] \le \operatorname{Var}[\phi(C)] \cdot \frac{1}{\alpha},$$

where we tested different values for $\alpha \in \mathbb{R}$.

An important aspect of the theory is the assumption that ϕ is normalized in the data generation process. Since this can not always be guaranteed in realworld problems, an approximation is necessary. Due to the small error, ϕ can be approximately normalized by normalizing E between 0 and 1. Since effect and cause variable are unknown beforehand and to ensure the same scaling, the cause data are also normalized.

5.1 Artificial data sets

We generated different artificial data sets which show the advantage of our approach over the existing mentioned techniques in different settings. In order to show that our approach is able to overcome the aforementioned drawbacks of existing methods, we tested different variances of the input and noise distribution. The generation process has the following form: $C \sim U(-b,b)$, $N_E \sim \mathcal{N}\left(0, \operatorname{Var}[\phi(C)] \cdot \frac{1}{\alpha}\right), E = \phi(C) + N_E$. We generated 100 data points for each evaluation and considered the functions $\phi(C) = C^2 + C, \phi(C) = C^3 + \sin(C)$ and $\phi(C) = \exp(C)$. Even though these functions are rather simple, they already point out the main differences between the approaches.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.



Fig. 1: Overview of the accuracies of different causal inference methods in artificial data sets. The figure caption indicates the underlying function ϕ , while the parameter α indicates the amount of noise; where the higher α is, the smaller the noise becomes. The parameter b implicitly controls the non-linearity of ϕ .

Figure 1 shows the results of evaluation with different artificial data sets. For all experiments, we varied the parameters α and b that determine the amount of noise and the degree of non-linearity of ϕ , respectively. Even though we focus on the setting with small noise, we also generated data with a high noise variance ($\alpha = 0.001$). In all evaluations, our algorithm has a clear advantage over existing methods and shows a high consistent accuracy. While ANM and LiNGAM perform poorly in settings with small noise variance ($\alpha = 1000$), IGCI generally struggles when b is small. The results for the high variance setting were the same for all three functions. Remarkably, our method also performed well in these very noisy data and even outperformed ANM and LiNGAM. Due to the space constraints, we omit additional information about the performance variances of the methods, but these are implicitly represented by the fluctuations in the accuracies.

5.2 Real-world data sets

For the evaluations in real-world data, we used the *CauseEffect* real-world benchmark data sets for causal inference taken from webdav.tuebingen.mpg.de/causeeffect/. These data sets consist of 95 various bivariate cause and effect pairs where the true causal direction is known beforehand.

These data sets have a generally high noise variance and may violate the made assumptions. Thus, an evaluation of the algorithm in these data shows the applicability to more general problems and also shows the performance in data with a high noise variance. For the evaluations, we considered a smoothing spline as a flexible regression function and the logistic function as a rather simple function. The results are shown in Table 1. Surprisingly, the less complex

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Algorithm	Accuracy
Our (smoothing spline)	$65.5\% \pm 2.03\%$
Our (logistic function)	$68.91\% \pm 2.63\%$
ANM	$65.9\% \pm 4.31\%$
IGCI	$69.15\% \pm 2.19\%$
LiNGAM	$48.52\% \pm 3.83\%$

Table 1: A comparison of different causal inference methods in the real-world data sets.

logistic function gives slightly better results than the complex smoothing spline function. Our algorithm may not outperform IGCI and ANM, but can compete with them and provide an alternative that can be easily implemented. The poor performance of LiNGAM can be explained by the non-linear data which violate LiNGAM's core assumption of linear relationships.

6 Conclusion

In this paper, we presented a novel causal inference principle for data with a small error variance. Our evaluations showed that this principle outperforms common causal inference approaches in data with small error noise. Even though there is no statement possible about data with a high noise variance, our method performed surprisingly well in big noise data and is able to compete with other techniques. Further advantages over many of the existing methods are the low requirement of qualitative data samples and and its easy implementation. In future work, we aim to extend the approach to the multivariate case and want to provide a confidence measure based on the error variance.

Acknowledgement

This work was supported by JSPS KAKENHI #25240036 and #24700275 and the Center of Innovation Program from the Japan Science and Technology Agency, JST.

References

- P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.
- [2] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.
- [3] P. Blöbaum, T. Washio, and S. Shimizu. Error asymmetry in causal and anticausal regression. arXiv preprint arXiv:1610.03263.
- [4] D. Janzing and B. Schölkopf. Semi-supervised interpolation in an anticausal learning scenario. Journal of Machine Learning Research, 16:1923–1948, 2015.
- [5] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.