Scaling Causal Inference in Additive Noise Models

Karim Assaad, Emilie Devijver, Éric Gaussier, Ali Ait-Bachir

Scaling Causal Inference in Additive Noise Models

Karim Assaad KARIM.ASSAAD@UNIV-GRENOBLE-ALPES.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Coservit
Grenoble, France

Emilie Devijver EMILIE.DEVIJVER@UNIV-GRENOBLE-ALPES.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
Grenoble, France

Eric Gaussier ERIC.GAUSSIER@IMAG.FR
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
Grenoble, France

Ali Aït-Bachir A.AIT-BACHIR@COSERVIT.COM
Coservit
Grenoble, France

Editor: Thuc Duy Le, Jiuyong Li, Kun Zhang, Emre Kıcıman, Peng Cui, and Aapo Hyvärinen

Abstract

The discovery of causal relationships from observations is a fundamental and difficult problem. We address it in the context of Additive Noise Models, and show, through both consistency analysis and experiments, that the state-of-art causal inference procedure on such models can be made simpler and faster, without loss of performance. Indeed, the method we propose uses one regressor instead of two in the bivariate case and $2(d-1)$ regressors instead of (d^2-1) in the multivariate case with d random variables. In addition, we show how one can, from the regressors we use, accelerate the computation of the Hilbert-Schmidt Independence Criterion, a standard independence measure used in several causal inference procedures.
Keywords: Causal Discovery, Additive Noise Models, Time Complexity, Autoencoders

1. Introduction

Causal inference has been the subject of many studies, Spirtes et al. (2000); Pearl (2000); Shimizu et al. (2006); Mooij et al. (2009); Zhang and Hyvärinen (2009, 2010); Bühlmann et al. (2014); Spirtes and Zhang (2016); Blöbaum et al. (2018) to name but a few. An important class of models to study causal inference are the so-called Additive Noise Models (ANMs), that simply consider, in the bivariate case, that the effect is a function of its cause *plus* a noise term independent of the cause. No further assumption is made regarding the function relating the cause to the effect. The corresponding structural causal model is given by:

$$\begin{aligned} C &:= N_c \\ E &:= f_E(C) + N_E, \qquad C \perp\!\!\!\perp N_E. \end{aligned}$$

An important property of ANMs is that they are usually identifiable except for some specific distributions contained in a 3-dimensional affine space (Hoyer et al., 2009). Within ANMs, the current best procedure to infer the causal structure of a set of variables (see for example the comparisons presented in Mooij et al. (2016) and Blöbaum et al. (2018)) is the one described in Mooij et al. (2009), which we refer to here as ANM-pHSIC. In this procedure, the direction of the causal relation is determined according to the lowest dependence between the potential cause and its residual when predicting the potential effect. The dependence is measured by the p-value of the empirical Hilbert-Schmidt Independence Criterion (HSIC) estimator (Gretton et al., 2005). For multivariate data sets, ANM-pHSIC relies on two main steps:

1. *Causal ordering* that consists in constructing a causal graph based on an ordering of the variables;
2. *Pruning* that consists in pruning the relations obtained in the causal ordering step.

However, ANM-pHSIC suffers from two main drawbacks:

1. Its reliance on the p-value of HSIC, and not directly on HSIC, is not well grounded theoretically;
2. Its time complexity limits its use to small to medium scale settings. Indeed ANM-pHSIC computes regression functions between all pairs of variables, which is of course problematic when the number of variables is important but also when the number of observations is important as each regression function will take more time to be estimated in this case. In addition, computing HSIC is a time consuming operation.

We specifically address these problems in this study and introduce a procedure that dispenses with training many regression functions. Intuitively, one can use an autoencoder to estimate the relations between all variables and *mask* (in a sense described below) some of the inputs and outputs of this autoencoder to obtain regressors between subsets of variables. By doing so, one dispenses with computing many different regressors. In addition, the regressors obtained are simple and scale well wrt the number of variables and observations. Lastly, the latent representations provided by the autoencoder can be used to accelerate the computation of HSIC.

The remainder of the paper is organized as follows: Section 2 provides a theoretical justification for the new procedure we propose, which is fully described in Section 3. Section 4 then presents a variety of experiments for the bivariate and multivariate cases. These experiments show that the new procedure we propose is indeed faster than ANM-pHSIC and leads to the same quality in terms of causal inference. Section 5 discusses our approach with respect to other studies and concludes the paper.

2. Considerations on the consistency of the causal ordering procedure in the bivariate case

Let assume two bivariate data sets, $\mathcal{D}_n := (x_i, y_i)_{i=1}^n$, and $\mathcal{D}_n^{'} := (x_i', y_i')_{i=1}^n$, both consisting of i.i.d. observations from $P_{X,Y}$ and let $\underline{x}$ denote the set of values $(x_1, \cdots, x_n)$ ($\underline{y}$, $\underline{x'}$,

... are defined in the same way). The **causal ordering procedure** (Mooij et al., 2016) for identifying bivariate causal graphs in ANMs can be summarized as follows:

1. Using $\mathcal{D}_n$, learn $\hat{f}_Y$ (resp. $\hat{f}_X$), an estimator of the regression function which maps x (resp. y) to $\mathbb{E}(Y|X=x)$ (resp. $\mathbb{E}(X|Y=y)$);
2. On $\mathcal{D}_n^{'}$, compute residuals $\hat{\underline{e}}'_Y = \underline{y}' - \hat{f}_Y(\underline{x}')$ and $\hat{\underline{e}}'_X = \underline{x}' - \hat{f}_X(\underline{y}')$;
3. Output $X \to Y$ if $\hat{C}(\underline{x}', \hat{\underline{e}}'_Y) < \hat{C}(\underline{y}', \hat{\underline{e}}'_X)$ and $Y \to X$ if $\hat{C}(\underline{y}', \hat{\underline{e}}'_X) < \hat{C}(\underline{x}', \hat{\underline{e}}'_Y)$, where $\hat{C}$ is an estimator of the dependence between the two variables (as measured through sets of values).

If the regression functions $\hat{f}_Y$ and $\hat{f}_X$ are *suitable* (i.e. the mean squared error between true and predicted residuals vanishes asymptotically in expectation) and if the score estimator $\hat{C}$ is consistent, then the above inference procedure is consistent.

As mentioned before, we want to use an autoencoder to estimate the relations between variables and then mask some of its inputs and outputs to obtain regressors between subsets of variables. The autoencoders we consider in this study are based on Multilayer Perceptrons (MLP) with only one hidden layer. Assuming a linear function at the output layer and a non-linear, squashing function σ at the input layer[1], the class of such MLPs takes the form:

$$\mathcal{F}_n = \left\{ \sum_{i=1}^{k_n} c_{i,j}\sigma(\mathbf{a_i}^T\mathbf{u} + b_i) + c_{0,j} : 1 \leq j \leq d', k_n \in \mathbb{N}, \right. \\ \left. (\mathbf{a_i}, \mathbf{u}) \in \mathbb{R}^d, b_i \in \mathbb{R}, \sum_{i=1}^{k_n}\sum_{j=1}^{d'} |c_{i,j}| \leq \beta_n \right\} \tag{1}$$

with d (resp. d'), k_n and β_n corresponding respectively to the dimension of the input (resp. output) of the MLP, to the number of hidden units and to a constraint on output weights. This class of function is weakly universally consistent:

Theorem 1 (extension of Theorem 16.1 of (Györfi et al., 2002) for $d' > 1$) *Let $\mathcal{F}_n$ be the class of neural networks defined in* (1)*, $\hat{f}_{mlp}(.;\mathcal{D}_n)$ be the network that minimizes the empirical L_2 risk in $\mathcal{F}_n$. If k_n and β_n satisfy, for $n \to +\infty$: $k_n \to +\infty, \beta_n \to +\infty$, and $k_n\beta_n^4 \log(k_n\beta_n^2)/n \to 0$, then $\hat{f}_{mlp}(.;.,\mathcal{D}_n)$ is weakly universally consistent for all distributions of input and output variables $(\mathbf{U}, \mathbf{V})$ with, for all $1 \leq j \leq d'$, $\mathbb{E}(V_j^2) < \infty$:*

$$\lim_{n\to\infty} \mathbb{E} \int ||\hat{f}_{mlp}(\mathbf{u};\mathcal{D}_n) - \mathbb{E}(\mathbf{V}|\mathbf{U}=\mathbf{u})||_2^2 d\mathbf{u} = 0.$$

Therefore, by Lemma 19 of (Mooij et al., 2016), $\mathbf{u} \mapsto \hat{f}_{mlp}(\mathbf{u};\mathcal{D}_n)$ is a suitable function. Let us now consider the case where $\mathbf{U} = \mathbf{V} = (X\ Y)^T$ and where the MLP considered is a *denoising* autoencoder (Vincent et al., 2008) that will be denoted by $\hat{f}_{ae}(.;\mathcal{D}_n)$. In our denoising autoencoder, one variable, randomly chosen, is arbitrarily set to 0 in the input, but not in the output, at each iteration during training, which enables to reconstruct a

1. In practice, we consider a more general class of functions.

corrupted version of the data. One thus considers different types of inputs, corresponding to whether or not a variable has been set to 0. We further denote by $\hat{f}^Y_{ae}$ (resp. $\hat{f}^X_{ae}$) the value predicted by the autoencoder for the output corresponding to Y (resp. X). Then, from Theorem 1, as all expectations are positive, one has:

$$\lim_{n\to\infty} \mathbb{E} \int (\hat{f}^Y_{ae}(\mathbf{u}; \mathcal{D}_n) - \mathbb{E}(\mathbf{Y}|\mathbf{U} = \mathbf{u}))^2 d\mathbf{u} = 0, \tag{2}$$

and similarly for $\hat{f}^X_{ae}$.
Focusing first on variable Y, we denote by $\mathbf{u}_{|y=0}$ the situation in which the input variable Y has been set to 0 and by $\mathbf{u}_{|y\neq 0}$ the situation in which it has not been changed. One can decompose the expectation in Eq. 2 according to these two cases:

$$\begin{aligned}\int (\hat{f}^Y_{ae}(\mathbf{u}; \mathcal{D}_n) - \mathbb{E}(\mathbf{Y}|\mathbf{U} = \mathbf{u}))^2 d\mathbf{u} = &\int (\hat{f}^Y_{ae}(\mathbf{u}_{|y=0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{Y}|\mathbf{U} = \mathbf{u}_{|y=0}))^2 d\mathbf{u}_{|y=0} \\ &+ \int (\hat{f}^Y_{ae}(\mathbf{u}_{|y\neq 0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{Y}|\mathbf{U} = \mathbf{u}_{|y\neq 0}))^2 d\mathbf{u}_{|y\neq 0}.\end{aligned} \tag{3}$$

Hence, exploiting again the fact that all quantities are positive in the right-hand side of Eq. (3) and that the left-hand side of Eq. (2) is equal to zero for $n \to \infty$, one obtains:

$$\lim_{n\to\infty} \mathbb{E} \int (\hat{f}^Y_{ae}(\mathbf{u}_{|y=0}; \mathcal{D}_n) - \mathbb{E}(\mathbf{Y}|\mathbf{U} = \mathbf{u}_{|y=0}))^2 d\mathbf{u}_{|y=0} = 0,$$

and similarly for $\hat{f}^X_{ae}$ and $\mathbf{u}_{|x=0}$.

Thus, the function $\mathbf{u} \mapsto \hat{f}^Y_{ae}(\mathbf{u}_{|y=0}; \mathcal{D}_n)$, regressing Y on X and obtained by setting the input Y of the denoising autoencoder considered above to 0, is weakly universally consistent. By Lemma 19 of Mooij et al. (2016), this function is also suitable, and so is the function $\mathbf{u} \mapsto \hat{f}^X_{ae}(\mathbf{u}_{|x=0}; \mathcal{D}_n)$ regressing X on Y.

Following Hoyer et al. (2009), we rely in this study on the Hilbert-Schmidt Independence Criterion (HSIC) for testing the independence of the estimated residuals with the input. Its empirical estimate, which is used as the score estimator $\hat{C}$, takes the form (Gretton et al., 2005)[2]:

$$\widehat{HSIC_k(\underline{x}, \underline{y})} := \frac{1}{(n-1)^2} tr(K_{\underline{x}} H K_{\underline{y}} H),$$

where $K_{\underline{x}}$ (resp. $K_{\underline{y}}$) is a $n \times n$ kernel matrix for $\underline{x}$ (resp. $\underline{y}$), and H is a $n \times n$ matrix defined by $H_{ij} = \delta_{ij} - 1$ for $1 \leq i, j \leq n$, where δ is the Kronecker symbol. Gretton et al. (2005) show that, when $n \to \infty$, HSIC $\to 0$ if and only if $X \perp\!\!\!\perp Y$. Furthermore, as shown in Mooij et al. (2016), the empirical HSIC estimator is consistent in the sense that as $n \to \infty$:

$$\widehat{HSIC_k(\underline{x}, \underline{y})} \xrightarrow{\text{P}} HSIC_k(\underline{x}, \underline{y}),$$

where k is a non-negative bounded kernel.

This leads us to the following consistency result[3]:

2. For simplicity, we restrict ourselves here to the symmetric version of HSIC where the same kernel is used for the two variables.
3. We conjecture that an extension of Theorem 2 exists in the case of more than two variables, but this is beyond the scope of this study.

Theorem 2 *Let X, Y be two real-valued random variables with joint distribution $P_{X,Y}$ that either satisfies an ANM $X \to Y$, or $Y \to X$, but not both. Suppose we are given a training data set $\mathcal{D}_n$ and a test data set $\mathcal{D}_n^{'}$ in the data splitting scenario. Let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a bounded non-negative Lipschitz-continuous kernel. Then, the* ***causal ordering procedure*** *in which $\hat{C} = \widehat{HSIC}_k$, $\hat{f}_Y(x) = \hat{f}_{ae}^{Y}(\mathbf{u}_{|y=0}; \mathcal{D}_n)$ and $\hat{f}_X(y) = \hat{f}_{ae}^{X}(\mathbf{u}_{|x=0}; \mathcal{D}_n)$ is a consistent procedure for estimating the direction of the ANM.*

The proof of Theorem 2 directly parallels the proof of Corollary 21 of (Mooij et al., 2016) and exploits the consistency of $\widehat{HSIC}$ and the suitability of the regression functions considered.

Based on these results, we propose in the next section a fast approach for multivariate causal discovery.

3. KIKO: a fast approach for causal discovery in ANM

Let us first extend the notations of the previous section to the multivariate case. We consider here d random variables that can be represented by a a random vector $\mathbf{X}$, the j^{th} coordinate of which is denoted by x_j; $\mathbf{X}_{\setminus j}$ denotes the random vector obtained from $\mathbf{X}$ by removing its j^{th} coordinate. The training and test data sets, $\mathcal{D}_n := (\mathbf{x}^{(i)})_{i=1}^n$ and $\mathcal{D}'_n := (\mathbf{x}'^{(i)})_{i=1}^n$, now consist of sets of vectors denoted by $\underline{\mathbf{x}}$. The denoising autoencoder learned on $\mathcal{D}_n$ and based on the single layer MLP is denoted by $\underline{\mathbf{x}} \mapsto \hat{f}_{ae}(\underline{\mathbf{x}}; \mathcal{D}_n)$, whereas $\underline{\mathbf{x}} \mapsto \hat{f}_{ae}^{x_j}(\underline{\mathbf{x}}_{|\underline{x}_j=0}; \mathcal{D}_n)$ denotes the function estimating x_j using the denoising autoencoder $\hat{f}_{ae}(\underline{\mathbf{x}}_{|\underline{x}_j=0}; \mathcal{D}_n)$ in which the j^{th} coordinate of all the vectors in $\underline{\mathbf{x}}$ is set to 0. The encoder part of the denoising autoencoder, in our case the function mapping an input to the single hidden layer, will be denoted by $\theta_{ae}(.)$; $\theta_{ae}(\underline{\mathbf{x}})$ corresponds to the application of the encoder to all the elements of the set $\underline{\mathbf{x}}$, leading to a set of latent representations of $\underline{\mathbf{x}}$. Finally, $\hat{f}_{mlp}(\underline{x}_j; \underline{\mathbf{x}}_{pa(j)}, \mathcal{D}_n)$ denotes the single layer MLP trained on $\mathcal{D}_n$ and predicting the n values $\underline{x}_j$ from $\underline{\mathbf{x}}_{pa(j)}$, whereas $\hat{f}_{mlp}(\underline{x}_j; \underline{\mathbf{x}}_{pa(j)|\underline{x}_l=0}, \mathcal{D}_n)$ denotes its restriction when setting the l^{th} variable to 0.

Our approach, called KIKO[4] and described in Algorithm 1, parallels ANM-pHSIC. It is decomposed into two parts. First, the variables are ordered according to their likelihood of being effects rather than causes. This is realized in the first loop of Algorithm 1 by estimating the variable for which the potential causes are less dependent to its residuals from the current set of variables. The empirical HSIC estimator is used to measure the independence between a latent-space representation of the input and residuals[5]. This procedure is a direct multivariate application of the results presented in Section 2. It starts by learning a complete autoencoder $\hat{f}_{ae}$ that is directly used, through a reminiscent mechanism of interventions, to derive a regression function $f_{ae}^{x_l}$ for each variable l from all the others. In practice, this is done either by cutting all the weights from the input variable x_l or by set-

4. *Knock-In Knock-Out*: one-for-one substitution (knock-in) of variables by zeros to decide which variable to knock out.
5. Note that here HSIC is used to measure the dependence between sequences of vectors and scalars, which raises no particular difficulties as kernels can be defined on both sequences.

ting the variable x_l to zero. This contrasts with ANM-pHSIC which needs to train specific regression functions for each variable.

Algorithm 1: KIKO: Causal Discovery Algorithm

Input $\mathcal{D}_n, \mathcal{D}'_n$
Identification of the causal ordering
for $j = d$ **downto** 1 **do**
  Learn $\hat{f}_{ae}(\mathbf{x}; \mathcal{D}_n)$
  For $l = 1$ **to** d **do**
    $\hat{\underline{e}}'_l = \underline{x}'_l - \hat{f}^{\mathrm{X}_l}_{ae}(\underline{\mathbf{x}}'_{|\underline{x}_l=0}; \mathcal{D}'_n)$
    $h_l = \widehat{HSIC}_k(\theta_{ae}(\underline{\mathbf{x}}'_{|\underline{x}_l=0}), \hat{\underline{e}}'_l)$
  end for
  $a_j = \arg\min_l h_l$
  $\mathbf{x} = \mathbf{x}_{\setminus a_j}$
end for
Pruning of spurious relations
for $j = 1$ **to** d **do**
  $pa(a_j) = \{a_1, \cdots, a_{j-1}\}$
  Learn $\hat{f}_{mlp}(x_{a_j}; \mathbf{x}_{pa(a_j)}, \mathcal{D}_n)$
  for l **in** $pa(a_j)$ **do**
    $\hat{\underline{e}}'_l = \underline{x}'_{a_j} - \hat{f}_{mlp}(\underline{x}_{a_j}; \underline{\mathbf{x}}_{pa(a_j)|\underline{x}_l=0}, \mathcal{D}_n)$
    $h_l = \widehat{HSIC}_k(\theta_{ae}(\underline{\mathbf{x}}_{pa(a_j)|\underline{x}_l=0}), \hat{\underline{e}}'_l)$
  end for
  $\alpha = GetGap(Sort(h))$
  $Prune(pa(a_j), \alpha)$
end for

After the first part, a causal graph that is likely to be faithful (*i.e.*, to contain all potential causal relations) is obtained and that still needs to be pruned to remove spurious causal relations. The second part aims at pruning these spurious relations between possible causes and effects identified in the first part. It does so by considering to which extent all possible causes of a variable are indeed independent from their residuals. So now we directly learn a regression function to each child using its potential parents. Then each parent iteratively undergo an intervention in order to check the significance of its influence on the child. The regression function used in this case is a single layer MLP $\hat{f}_{mlp}$, with the set of possible causes as input and the effect considered as output. The empirical HSIC estimator is again used to measure dependency. The function *GetGap* selects the first gap between consecutive HSIC scores with a considerable increase (2 times) with respect to its predecessor. All relations corresponding to an HSIC smaller to this gap are pruned.

Remark on the computation of $\widehat{HSIC}_k$ Autoencoders compress their input into a latent representation and then reconstruct the output from this representation. The latent representation can be seen as a noise-free version of the input, that contains all its important information (hence the fact that one can *reconstruct* the input from this representation). The latent representation can thus be used as a proxy to the input in the computation of

$\widehat{HSIC}_k$. By doing so, one relies on a lower dimensional version of the input that can lead to significant gains when computing the Gram matrices used in $\widehat{HSIC}$. Assuming that the complexity for computing the kernel between two observations is $\mathcal{O}(d)$, then the gain in complexity by using a latent representation of dimension d'' will be d/d''. As we will see in the next section, this gain can indeed be substantial.

4. Experiments

As mentioned before, the main objective of this study is to propose a method for causal inference in ANMs that scales well while providing the same level of quality as ANM-pHSIC. The experiments[6] below illustrate these two points.

4.1 Quality of causal inference

We make use here of the commonly used Cause-Effect Pairs (CEP) benchmark of 100 real-world data sets, introduced and available in Mooij et al. (2016); Dheeru and K. Taniskidou (2017), and an artificial multivariate data set introduced in Hoyer et al. (2009) with 500 observations and 20 variables arranged in the diamond-like causal structure. As mentioned before, we place ourselves in the data splitting scenario: the main data set is split into two equal parts, the first part being used to fit the data and the second one to estimate the residuals and compute the HSIC[7] score. In addition, another scenario, called data recycling, is used for the concurrent method, denoted by ANM-pHSIC(R), where $\mathcal{D}_n = \mathcal{D}'_n$. Note that for ANM-pHSIC methods, the p-value is computed using the γ approximation, which speeds up the computation.

For the Gram matrices used in $\widehat{HSIC}_k$, two RBF kernels are used with bandwidths selected by the median heuristics (Schölkopf and Smola, 2001). In all neural networks, we consider one hidden layer with 10 neurons. Adam optimizer is used with a learning rate 0.01 and 300 epochs. All observed variables are scaled into $[-1, 1]$. The denoising autoencoder is set to denoise an observation with a probability 0.5. In case of denoising, one chooses one variable at random and forces its value to 0, while the others are left untouched. This means that, for bivariate datasets, around one fourth of the training examples have a null value for each variable. We consider two different architectures[8] suited with the theory presented in Section 2, the first relies on a Tanh at the input layer and a linear function in the output layer, and the second relies on a Leaky Relu at the input layer and a Tanh at the output layer, which are donoted respectively as KIKO-TL and KIKO-RT. In addition, we also consider a simplified variant of the above in which the denoising autoencoder is replaced by a standard autoencoder. These two variants are denoted as SKIKO-TL and SKIKO-RT.

Evaluation measures. To evaluate the quality of the inferred graph, we rely on different measures, depending on the number of variables. For the bivariate case, we use the accuracy measure ACC. For the multivariate case, one has two aspects to consider, namely

6. The code is available at `https://github.com/kassaad/KIKO/`.
7. We compute HSIC by expanding $tr(K_{\underline{x}} H K_{\underline{y}} H)$ into non-repeated terms (Gretton et al., 2005).
8. We investigated more complex architectures (the only constraint is that the first hidden layer does not contain sharing weights) but surprisingly, the simplest performs the best.

the identification of the cause-effect order and the pruning of the spurious relations. We use a stability measure STAB for the former, that is 1 if the order is correctly predicted and 0 otherwise, and a similarity measure SIM based on the Hamming distance for the latter. This similarity is defined as:

$$\text{SIM} = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathbb{1}_{P_{i,j}=T_{i,j}},$$

where P and T are the adjacency matrices associated to the constructed DAG and the ground truth DAG respectively. All those measures are between 0 and 1, 1 being the best. The results of each method are averaged over 100 runs so that one can estimate the mean and the variance of both accuracy and similarity.

Numerical results. As shown in Table 1 in the bivariate case, the four variants of our method (KIKO-TL, KIKO-RT, SKIKO-TL and SKIKO-RT) yield similar results. Whereas our main goal was to speed up the procedure, KIKO outperforms ANM-pHSIC and achieves similar results as ANM-pHSIC(R) while being computationally faster. Note that, for all methods, the accuracy has small variance. In the multivariate case, one can note that KIKO-based methods are discovering the true ordering, as the stability over the 100 runs with the true DAG is better than ANM-pHSIC and ANM-pHSIC(R). In the pruning phase, KIKO-RT and SKIKO-RT yield slightly better results. The fact that SKIKO yields slightly better results than KIKO on the bivariate datasets may be due to denoising that may lead to a too aggressive strategy when the number of variables is small. This is a point we plan to investigate in the future.

Table 1: Results averaged over 100 runs on the Cause-Effect Pairs and the simulated data sets, in terms of accuracy (ACC) on the 100 bivariate data sets and stability (STAB) and similarity (SIM) on the simulated data set. Standard deviations are given in parenthesis and the best results are in bold.

	CEP	Simulated data	
Method	ACC	STAB	SIM
KIKO-RT	62.6% (0.02)	89%	**0.86 (0.09)**
KIKO-TL	61.2% (0.02)	**100%**	0.77 (0.08)
SKIKO-RT	**64.5% (0.02)**	**100%**	**0.86 (0.12)**
SKIKO-TL	63.9% (0.02)	**100%**	0.79 (0.07)
ANM-pHSIC	57.0% (0.03)	63%	0.82 (0.09)
ANM-pHSIC(R)	62.5% (0.01)	88%	0.83 (0.04)

4.2 Time complexity

To illustrate the gain in time we use simulated data[9] in two different settings. In the first one, we neglect the calculation of $\widehat{HSIC}_k$, which will allow us to demonstrate the advan-

9. This simulated data is obtained from normal distributions and can be generated with the code available at `https://github.com/kassaad/KIKO/`.

tage of learning $2(d-1)$ models, instead of (d^2-1), by learning one global model, which is then used to derive specialized regression functions. In the second one, we take into account the calculation of $\widehat{HSIC}_k$ and show the advantage of relying on latent representations provided by autoencoders.

Disregarding (for the moment) the computation of $\widehat{HSIC}_k$, the time complexity of KIKO[10] is $\mathcal{O}(2(d-1)M_{mlp} + (d^2-1)P_{mlp})$, where M_{mlp} and P_{mlp} corresponds respectively to the time complexity of training a single layer MLP and of predicting values using a single layer MLP. In contrast, the time complexity of ANM-pHSIC is $\mathcal{O}(M_{gp}P_{gp}(d^2-1))$, where M_{gp} and P_{gp} corresponds respectively to the time complexity of training a Gaussian process and of predicting values using a Gaussian process. To illustrate this difference, we first show how the computation times of ANM-pHSIC and of KIKO evolve wrt the number of variables and the number of observations while disregarding the computation of $\widehat{HSIC}_k$ (Figure 1). As one can note, KIKO outperforms ANM-pHSIC in both cases. For example, in the case of 80 variables and 3,000 observations, KIKO is around 6 times faster than ANM-pHSIC. For 20 variables and 10,000 observations, KIKO is around 5 times faster than ANM-pHSIC. This gain can be explained by the fact that KIKO relies on less models, these models being furthermore simpler and faster to estimate.

To illustrate the advantage of using latent-space representation in $\widehat{HSIC}_k$, we have repeated these experiments without disregarding the calculation of $\widehat{HSIC}_k$ (Figure 2). For the sake of fairness, we calculate in both algorithms $\widehat{HSIC}_k$ and its p-value (although in KIKO we do not need the p-value). So the only difference between KIKO and ANM-pHSIC concerning $\widehat{HSIC}_k$ is that the former uses the latent representations provided by the autoencoders while the latter directly relies on the input variables. As one can note from Figure 2, here again KIKO is significantly faster than ANM-pHSIC (around 6 times faster for 80 variables and 3,000 observations, and around 2 times faster for 20 variables and 10,000 observations). When the number n of observations is important, the computation of $\widehat{HSIC}_k$, quadratic in n, becomes the limiting factor. However, due to the use of latent representations, KIKO can still be used for large-scale data. Lastly, note that ANM-pHSIC(R) uses more data and is even slower than the version of ANM-pHSIC used in our experiments.

5. Discussion and Conclusion

Several families of methods have been developed to infer causal relations from observational data. In this paper we focused on ANMs, a class belonging to the family that uses causal footprints. ANMs rely on the assumption that the noise is additive and independent of the cause. In the literature other methods can be found within the same family. LiNGAM (Linear Non-Gaussian Acyclic Model, Shimizu et al. (2006)) is an ANM that is restricted to linear transformations and non-Gaussianity. CAM (Causal Additive Model, Bühlmann et al. (2014)) is a restricted class of ANMs suited for high-dimensional data that assumes Gaussian errors and uses maximum likelihood principle.

Generalization of additive noise models have also been introduced. PNL (Post Non Linear models, Zhang and Hyvärinen (2009)) is a generalization of ANM that allows for a

10. KIKO-based methods have the same time complexity, as only the activation functions are changing.

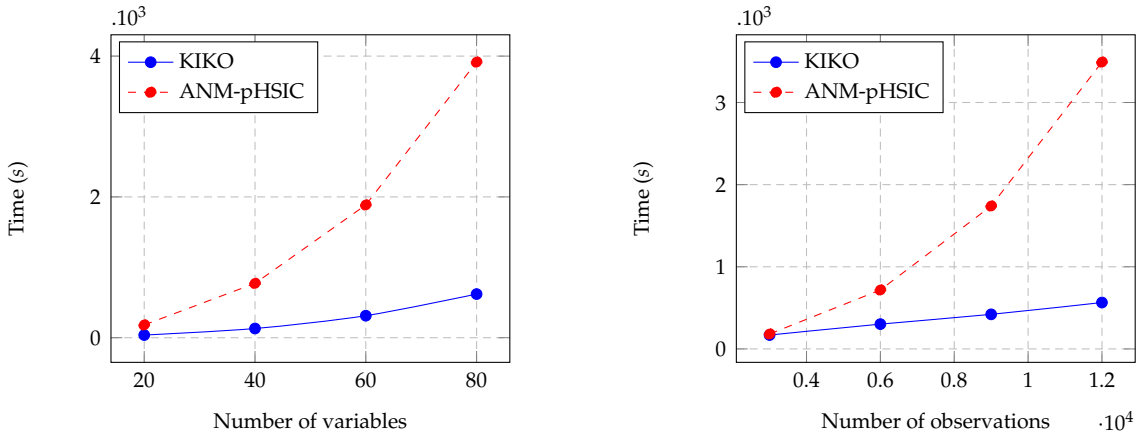

Figure 1: Time complexity of KIKO vs ANM-pHSIC by: (Left) number of variables (sample size fixed to 3000); (Right) sample size (number of variables fixed to 20), where the independence measure $\hat{C}$ is considered as an atomic operation.

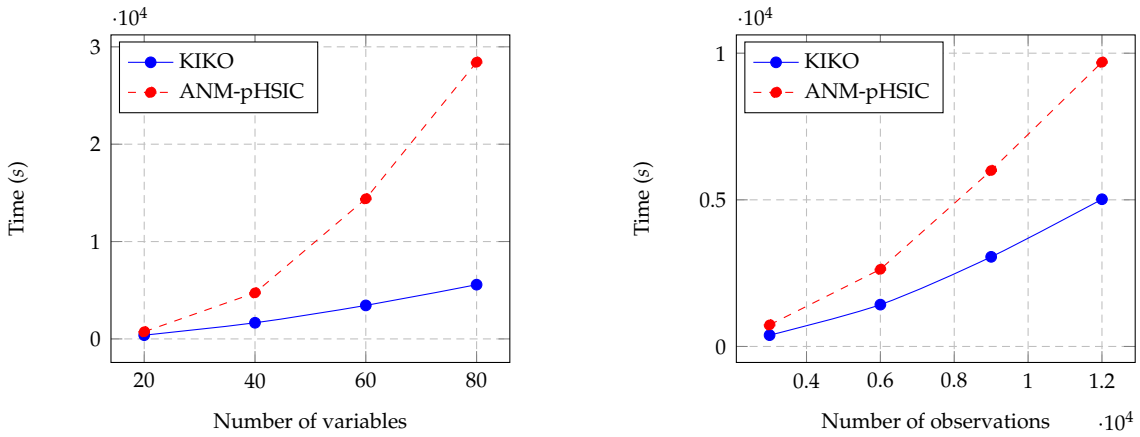

Figure 2: Time complexity of KIKO vs ANM-pHSIC by: (Left) number of variables (sample size fixed to 3000); (Right) sample size (number of variables fixed to 20), where the independence measure $\hat{C}$ is HSIC.

post-nonlinear transformation of the variables. Different viewpoints have also been considered. IGCI (Information-Geometric Causal Inference, Janzing et al. (2012)) illustrates a different class of methods, without noise, and determines the causal relations under a different independence assumption. CURE (Causal inference with Unsupervised inverse REgression, Sgouritsa et al. (2015)) is based on the idea that the distribution of the cause does not help to infer the effect, contrarily to the distribution of the effect which does help to infer the cause. RCC (Randomized Causation Coefficient, Lopez-Paz et al. (2015)) phrases causal inference as a supervised learning problem in a kernel mean embeddings frame-

work. RECI (Regression Error based Causal Inference, Blöbaum et al. (2018)) is based on an asymmetry in the prediction error and allows a dependency between cause and noise.

Within the class of additive noise models, ANM-pHSIC is known to be the best performing method (Blöbaum et al., 2018). Our goal here was to show that this algorithm could be made simpler and faster through the use of MLPs and intervention-like mechanisms. To do so, we introduced a procedure, which we refer to as KIKO, that uses autoencoders, which are then specialized to behave as regression functions. This specialization allows to dispense with the training of many regression functions, which explains, together with the use of MLPs in lieu of Gaussian regression processes and the use of the latent representations in the computation of HSIC, the gain obtained by our method, that can be used, unlike ANM-pHSIC, in large-scale settings when the number of variables and observations are important.

References

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 900–909. PMLR, 2018.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42(6):2526–2556, 2014.

Dua Dheeru and Efi K. Taniskidou. UCI machine learning repository, 2017.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, 2005.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.

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 21*. ACM Press, 2009.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182-183:1–31, 2012.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1452–1461, 2015.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th International Conference on Machine Learning*, pages 745–752, New York, NY, USA, 2009. Max-Planck-Gesellschaft, ACM Press.

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

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

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *JMLR Workshop and Conference Proceedings*, pages 847–855, 2015.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, Feb 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x. URL `https://doi.org/10.1186/s40535-016-0018-x`.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294. URL `http://doi.acm.org/10.1145/1390156.1390294`.

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*, UAI '09, pages 647–655, Arlington, Virginia, United States, 2009. AUAI Press.

Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf, editors, *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pages 157–164, Whistler, Canada, 12 Dec 2010. PMLR. URL `http://proceedings.mlr.press/v6/zhang10a.html`.