Inference of Cause and Effect with Unsupervised Inverse Regression

Eleni Sgouritsa **Dominik Janzing** **Philipp Hennig** **Bernhard Schölkopf**
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{eleni.sgouritsa, dominik.janzing, philipp.hennig, bs}@tuebingen.mpg.de

Abstract

We address the problem of causal discovery in the two-variable case given a sample from their joint distribution. The proposed method is based on a known assumption that, if $X \to Y$ (X causes Y), the marginal distribution of the cause, $P(X)$, contains no information about the conditional distribution $P(Y|X)$. Consequently, estimating $P(Y|X)$ from $P(X)$ should not be possible. However, estimating $P(X|Y)$ based on $P(Y)$ may be possible.

This paper employs this asymmetry to propose CURE, a causal discovery method which decides upon the causal direction by comparing the accuracy of the estimations of $P(Y|X)$ and $P(X|Y)$. To this end, we propose a method for estimating a conditional from samples of the corresponding marginal, which we call unsupervised inverse GP regression. We evaluate CURE on synthetic and real data. On the latter, our method outperforms existing causal inference methods.

1 INTRODUCTION

Drawing causal conclusions for a set of observed variables given a sample from their joint distribution is a fundamental problem [Spirtes et al., 2000, Pearl, 2009]. Our goal here is to decide between $X \to Y$ and $Y \to X$ (assuming no latent confounders) for two continuous univariate random variables X and Y, given a sample from their joint distribution, $P(X,Y)$. Conditional-independence-based causal discovery methods [Spirtes et al., 2000, Pearl, 2009] estimate the *Markov equivalent* graphs, all entailing the same conditional independences. However, in the case of only two variables, these methods can not recover the causal graph based on $P(X,Y)$, since $X \to Y$ and $Y \to X$ are Markov equivalent.

We review various methods that are also appropriate for the case of two variables. Hoyer et al. [2009] and Peters et al. [2014] suggest using Additive Noise Models (ANM). The causal inference method then reads: whenever $P(X,Y)$ allows for an ANM in one direction, i.e., there is a function f and a noise variable E such that $Y = f(X) + E$ with $E \perp\!\!\!\perp X$, but not in the other, i.e., X cannot be obtained as a function of Y plus independent noise, then the former direction is inferred to be the causal one (in this case $X \to Y$). They further show that in the generic case (up to some exceptions like the case of linear f and Gaussian X and E) the model is identifiable, that is, if there is an ANM in one direction, the joint distribution $P(X,Y)$ does not allow for an ANM in the backward direction. Previous work by Shimizu et al. [2006] proves identifiability of ANM when restricted to linear functions and non-Gaussian input and noise distributions (Linear Non-Gaussian Acyclic Model (LiNGAM)). A generalization of ANM is the Post-Nonlinear Model (PNL) [Zhang and Hyvärinen, 2009], where $Y = h(f(X) + E)$, with $E \perp\!\!\!\perp X$, which is also identifiable, except for some special cases. Mooij et al. [2010] infer the causal direction by Bayesian model selection, defining non-parametric priors on the distribution of the cause and the conditional of the effect given the cause. Other causal inference methods are based on the following postulate [Janzing and Schölkopf, 2010, Janzing et al., 2012, Daniusis et al., 2010, Schölkopf et al.]:

Postulate 1 (indep. of input and mechanism) *If $X \to Y$, then the marginal distribution of the cause, $P(X)$, and the conditional distribution of the effect given the cause, $P(Y|X)$, are "independent" in the sense that $P(Y|X)$ contains no information about $P(X)$ and vice versa.*

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

The (causal) conditional $P(Y|X)$ can be thought of as the *mechanism* transforming cause X to effect Y. Then, Postulate 1 is plausible if we are dealing with a mechanism of nature that does not care what (input $P(X)$) we feed into it. This independence can be violated in the backward direction: $P(Y)$ and $P(X|Y)$ may contain information about each other because each of them inherits properties from both $P(X)$ and $P(Y|X)$. This constitutes an asymmetry between cause and effect. While Postulate 1 is abstract, the aforementioned approaches provide formalizations by specifying what is meant by *independence* or *information*: Janzing and Schölkopf [2010] postulate *algorithmic* independence of $P(Y|X)$ and $P(X)$, i.e. zero algorithmic mutual information: $I(P(X) : P(Y|X)) \stackrel{+}{=} 0$. This is equivalent to saying that the shortest description (in the sense of Kolmogorov complexity) of $P(X, Y)$ is given by separate descriptions $P(X)$ and $P(Y|X)$. Since Kolmogorov complexity is uncomputable, practical implementations must rely on other notions of (in)dependence or information. For deterministic non-linear relations $Y = f(X)$, Janzing et al. [2012] and Daniusis et al. [2010] define independence through uncorrelatedness between $\log f'$ and the density of $P(X)$, both viewed as random variables (note that in this case $P(Y|X)$ is completely determined by f). This is reformulated in terms of information geometry as a certain orthogonality in information space. The corresponding causal inference method sometimes also works for sufficiently small noise. Finally, Schölkopf et al. do not propose a new causal inference method but argue that *knowing* the causal direction has implications for various learning scenarios, including semi-supervised learning (SSL). Specifically, if $X \to Y$, $P(X)$ contains no information about $P(Y|X)$ according to Postulate 1. As a result, a more accurate estimate of $P(X)$, as may be possible by the addition of the extra unlabeled points in SSL, does not influence an estimate of $P(Y|X)$, and SSL is pointless in this scenario. In contrast, SSL may be helpful in case $Y \to X$. Thus, their notion of independence between $P(X)$ and $P(Y|X)$ implicitly reads: the former is not helpful for estimating the latter.

The proposed method is inspired by the latter. Our use of Postulate 1 complies with their notion of independence: if $X \to Y$, estimating $P(Y|X)$ based on $P(X)$ should not be possible. In contrast, estimating $P(X|Y)$ given $P(Y)$ may be possible. Employing this asymmetry, we propose CURE, a method to infer the causal graph in the case of two variables that is appropriate for non-deterministic relations. The proposed causal inference method infers $X \to Y$ if the estimation of $P(X|Y)$ based on $P(Y)$ is more accurate than the one of $P(Y|X)$ based on $P(X)$. Otherwise, $Y \to X$ is inferred.

To this end, we propose a method for estimating a conditional distribution based on samples from the corresponding marginal. We call it unsupervised inverse GP regression for the following reason: in standard supervised regression, given a sample from $P(X, Y)$, the goal is to estimate the conditional $P(Y|X)$. We call supervised *inverse* regression the task of estimating the conditional $P(X|Y)$, without changing the original regression model of Y on X that was used for the estimation of $P(Y|X)$. The reason for introducing inverse regression is related to our goal of estimating the conditional $P(X|Y)$ from samples of the marginal $P(Y)$. Using standard regression of X on Y would be pointless in this scenario since only samples from $P(Y)$ are given. As a result, inverse regression is chosen and since it is based only on data from $P(Y)$ and not $P(X, Y)$, we call it *unsupervised inverse regression*. Finally, we term the proposed causal discovery method Causal inference with Unsupervised inverse REgression (CURE).

Sections 2 and 3.2 describe the building blocks for unsupervised inverse regression, presented in Section 3.1. Section 4 describes CURE. In the following, we denote random variables with capital letters and their corresponding values with lower case letters. Random vectors are denoted with bold face capital letters and their values with bold face lower case letters.

2 GAUSSIAN PROCESS LATENT VARIABLE MODEL

The Gaussian process latent variable model (GP-LVM) [Lawrence, 2005] can be interpreted as a multi-output Gaussian process (GP) model [Rasmussen and I., 2006] in which only the output data are observed, while the input remain latent. Let $\mathbf{y}^* \in \mathbb{R}^{N \times D}$ be the observed data where N is the number of observations and D the dimensionality of each observation. These data are associated with a latent vector taking values $\mathbf{x} \in \mathbb{R}^{N \times Q}$. The purpose is often dimensionality reduction, thus $Q \ll D$. GP-LVM defines a mapping from the latent to the observed space by using GPs, with hyperparameters $\boldsymbol{\theta}$. Assuming independence across the dimensions, the likelihood function is given as:

$$p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta}) = \prod_{d=1}^{D} p(\mathbf{y}^*_{\mathbf{d}}|\mathbf{x}, \boldsymbol{\theta})$$

where $\mathbf{y}^*_{\mathbf{d}}$ the d^{th} column of $\mathbf{y}^*$, $p(\mathbf{y}^*_{\mathbf{d}}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}^*_{\mathbf{d}}; \mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N)$, and $K_{\mathbf{x},\mathbf{x}}$ the $N \times N$ covariance function defined by a selected kernel function. Thus, $p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta})$ is a product of D independent Gaussian processes where the input, $\mathbf{x}$, is latent.

For the present work, only univariate random variables

are relevant, thus $D = Q = 1$. This defines a *single-output* GP-LVM, i.e., just one GP model with latent input. In this case, $\mathbf{y}^* \in \mathbb{R}^N$, $\mathbf{x} \in \mathbb{R}^N$ and the likelihood function of single-output GP-LVM is given as:

$$p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}^*; \mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N) \quad (1)$$

with $\boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$, where we choose the RBF kernel

$$k(x_i, x_j) = \{K_{\mathbf{x},\mathbf{x}}\}_{i,j} = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}(x_i - x_j)^2\right)$$

Lawrence [2005] finds $\mathbf{x}$ (for multiple-output GP-LVM), by MAP estimation, selecting a Gaussian prior for $\mathbf{x}$, while jointly maximizing with respect to $\boldsymbol{\theta}$. In Bayesian GP-LVM [Titsias and Lawrence, 2010], instead, $\mathbf{x}$ is variationally integrated out and a lower bound on the marginal likelihood $p(\mathbf{y}^*)$ is computed.

3 UNSUPERVISED INVERSE REGRESSION

Throughout the rest of the paper, let $\mathbf{x}^* := (x_1^*, \ldots, x_N^*)$ and $\mathbf{y}^* := (y_1^*, \ldots, y_N^*)$ be a sample of N independently and identically distributed (i.i.d.) observations from $P(X, Y)$. Moreover, $\mathbf{x}^*$ and $\mathbf{y}^*$ are rescaled between zero and one. In standard supervised regression, given $\mathbf{x}^*$ and $\mathbf{y}^*$, the task is to estimate the conditional distribution $P(Y|X)$, i.e., compute the predictive distribution $p(y|x, \mathbf{x}^*, \mathbf{y}^*)$. In supervised *inverse* regression (or simply inverse regression) the task is to obtain $p(x|y, \mathbf{y}^*, \mathbf{x}^*)$, but without changing the original regression model of Y on X. Finally, the task of *unsupervised inverse regression* is to compute $p(x|y, \mathbf{y}^*)$, i.e. estimate $P(X|Y)$ based *only* on samples $\mathbf{y}^*$ from $P(Y)$ (and not $\mathbf{x}^*$). In the following unsupervised inverse GP regression is described.

3.1 Unsupervised Inverse GP Regression

The goal is to compute $p(x|y, \mathbf{y}^*)$. A Gaussian process regression model of Y on X is used. The predictive distribution $p(x|y, \mathbf{y}^*)$ is given by marginalizing over the distribution of the latent random vector (N-dimensional latent variable) $\mathbf{X} := (X_1, \ldots, X_N)$ (instead of inserting the true values $\mathbf{X} = \mathbf{x}^*$) and the unknown GP hyperparameters $\boldsymbol{\Theta}$:

$$\begin{aligned} p(x|y, \mathbf{y}^*) &= \int_{\mathcal{X},\Theta} p(\mathbf{x}, \boldsymbol{\theta}, x|\mathbf{y}^*, y) d\mathbf{x} d\boldsymbol{\theta} \\ &= \int_{\mathcal{X},\Theta} p(x|y, \mathbf{y}^*, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*, y) d\mathbf{x} d\boldsymbol{\theta} \\ &\approx \int_{\mathcal{X},\Theta} p(x|y, \mathbf{y}^*, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*) d\mathbf{x} d\boldsymbol{\theta} \quad (2) \end{aligned}$$

The first factor, $p(x|y, \mathbf{y}^*, \mathbf{x}, \boldsymbol{\theta})$, is the predictive distribution of *supervised inverse* GP regression, which is explained in section 3.2 (Eq. (5)). The second factor, $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$, is the posterior distribution over $\mathbf{x}$ and the hyperparameters $\boldsymbol{\theta}$, given the observed $\mathbf{y}^*$.

A uniform prior, $\mathcal{U}(0, 1)$, is chosen for the unknown distribution of X. A uniform prior is, additionally, placed over $\boldsymbol{\theta}$ which suppresses overly flexible functions (small ℓ) to restrict the function class. By Bayes' theorem:

$$\begin{aligned} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*) &= \frac{p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}) p(\boldsymbol{\theta})}{p(\mathbf{y}^*)} \propto p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}) p(\boldsymbol{\theta}) \\ &= p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}^*; \mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N) \quad (3) \end{aligned}$$

$p(\mathbf{y}^*|\mathbf{x}, \boldsymbol{\theta})$ is the likelihood of single-output GP-LVM (Eq. (1)). Note that the computation of the latent's posterior distribution $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$ is analytically intractable since $\mathbf{x}$ appears non-linearly inside the inverse of $K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N$ [Titsias and Lawrence, 2010]. In our implementation, we approximate the posterior $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$ using a Markov Chain Monte Carlo (MCMC) method, slice sampling [Neal, 2003]. The sample size N determines the dimensionality of the space to sample from, which is $N + 3$ (including the three hyperparameters). Thus, the computational complexity is determined by N and this step poses the main computational bottleneck of our algorithm.

$p(x|y, \mathbf{y}^*)$ is estimated by replacing the integral in (2) with a sum over M MCMC samples from $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$:

$$p(x|y, \mathbf{y}^*) \approx \frac{1}{M} \sum_{i=1}^{M} p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i) \quad (4)$$

So, $p(x|y, \mathbf{y}^*)$ is computed as the average of predictive distributions of supervised inverse regressions. Each predictive distribution $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$ uses the i^{th} sample, $(\mathbf{x}^i, \boldsymbol{\theta}^i)$, from the posterior $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$.

3.2 Supervised Inverse GP Regression

Following from the previous section, the remaining task is to compute $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$ in (Eq. (4)), for each MCMC sample i, with $1 \leq i \leq M$. Since $\boldsymbol{\theta}^i$ and $\mathbf{x}^i$ are independent and the distribution of X is uniform, by Bayes' theorem we get:

$$\begin{aligned} p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i) &\propto p(\mathbf{y}^*, y|\mathbf{x}^i, x, \boldsymbol{\theta}^i) p(x|\mathbf{x}^i, \boldsymbol{\theta}^i) \\ &= \mathcal{N}(\mathbf{y}^*, y; \mathbf{0}, K_{(\mathbf{x}^i,x),(\mathbf{x}^i,x)} + {\sigma_n^i}^2 I_N) \quad (5) \end{aligned}$$

Notice that, unlike standard GP regression, the predictive distribution of inverse GP regression, $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$, is not Gaussian. We first compute $\mathcal{N}(\mathbf{y}^*, y; \mathbf{0}, K_{(\mathbf{x}^i,x),(\mathbf{x}^i,x)} + {\sigma_n^i}^2 I_N)$ at the points of a grid on $[0, 1]$, and then normalize appropriately to get $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$. Fig. 1 illustrates an example of supervised inverse regression. The predictive distributions of standard GP regression, $p(y|x, \mathbf{x}^i, \mathbf{y}^*, \boldsymbol{\theta}^i)$ (for

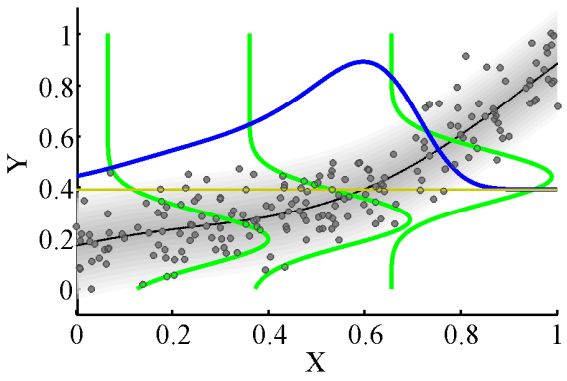

Figure 1: The predictive distributions of standard GP regression at three x values (green) and the predictive distribution of supervised inverse GP regression at one y value (blue).

some i), at three x values are depicted in green and the predictive distribution of inverse GP regression, $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$, at one y value (yellow line), in blue.

The usual practice to estimate $p(x|y, \mathbf{y}^*, \mathbf{x}^i)$ would be to learn directly a map from Y to X (discriminative model). However, we need to use GP regression of Y on X and not of X on Y in order to comply with the model used in Section 3.1.

To conclude, $p(x|y, \mathbf{y}^*)$ is computed from Eq. (4), using Eq. (5) for $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$. Likewise, we can compute $p(y|x, \mathbf{x}^*)$ repeating the above procedure with a GP regression model of X on Y.

3.3 Evaluation

Finally, we need to evaluate the accuracy of our estimation of $P(X|Y)$. We compute the negative log likelihood $L^{\text{unsup}}_{X|Y} = -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i^*|y_i^*, \mathbf{y}^*)$ at $\mathbf{x}^*$, $\mathbf{y}^*$ to measure the performance of unsupervised inverse regression. We could also evaluate it at new test points if we had a separate test set $\mathbf{x}^{\text{te}}$, $\mathbf{y}^{\text{te}}$ as $-\frac{1}{N}\sum_{i=1}^{N} \log p(x_i^{\text{te}}|y_i^{\text{te}}, \mathbf{y}^*)$. However, since the task is unsupervised, we don't have overfitting issues and use all data for estimating $P(X|Y)$. In order to evaluate the accuracy of the estimation of $P(X|Y)$, we compare $L^{\text{unsup}}_{X|Y}$ with the accuracy of the corresponding supervised inverse regression $L^{\text{sup}}_{X|Y} = -\frac{1}{N}\sum_{i=1}^{N} \log p(x_i^*|y_i^*, \mathbf{y}^*, \mathbf{x}^*)$, using again a uniform prior for X but with $\boldsymbol{\theta}$ computed by maximization of $p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$. This way, we measure how much the performance degrades due to the absence of $\mathbf{x}^*$, specifically by:

$$D_{X|Y} = L^{\text{unsup}}_{X|Y} - L^{\text{sup}}_{X|Y} \tag{6}$$

4 CURE

The ultimate goal is to decide upon the causal direction, $X \to Y$ or $Y \to X$, given $\mathbf{x}^*$ and $\mathbf{y}^*$. According to Postulate 1, if $X \to Y$, estimating $P(Y|X)$ from $P(X)$ should not be possible. In contrast, estimating $P(X|Y)$ based on $P(Y)$ may be possible. So, CURE is given as follows: if we can estimate $P(X|Y)$ based on samples from $P(Y)$ more accurately than $P(Y|X)$ based on samples from $P(X)$, then $X \to Y$ is inferred. Otherwise, $Y \to X$ is inferred. In particular, we apply unsupervised inverse GP regression two times. First, $D_{X|Y}$ is computed as in (6):

$$D_{X|Y} = L^{\text{unsup}}_{X|Y} - L^{\text{sup}}_{X|Y} =$$
$$-\frac{1}{N}\sum_{i=1}^{N} \log p(x_i^*|y_i^*, \mathbf{y}^*) + \frac{1}{N}\sum_{i=1}^{N} \log p(x_i^*|y_i^*, \mathbf{y}^*, \mathbf{x}^*)$$

to evaluate the estimation of $P(X|Y)$ based on $\mathbf{y}^*$. Then, $D_{Y|X}$ is computed as:

$$D_{Y|X} = L^{\text{unsup}}_{Y|X} - L^{\text{sup}}_{Y|X} =$$
$$-\frac{1}{N}\sum_{i=1}^{N} \log p(y_i^*|x_i^*, \mathbf{x}^*) + \frac{1}{N}\sum_{i=1}^{N} \log p(y_i^*|x_i^*, \mathbf{x}^*, \mathbf{y}^*)$$

to evaluate the estimation of $P(Y|X)$ based on $\mathbf{x}^*$. Finally, we compare the two performances: if $D_{X|Y} < D_{Y|X}$, then we infer the causal direction to be $X \to Y$, otherwise we output $Y \to X$.

5 DISCUSSION

Figure 2 depicts three datasets generated according to the causal model $X \to Y$ (grey points) (note the exchanged axes in the last figure). Since $X \to Y$ we expect to be able to estimate $P(X|Y)$ (based on $P(Y)$) more accurately than $P(Y|X)$ (based on $P(X)$). The quality of the estimation strongly depends on the generated MCMC samples from the high-dimensional posterior in (3). Figures 2(a) and 2(b) refer to the estimation of $P(X|Y)$ based on samples from $P(Y)$, whereas Fig. 2(c) to the estimation of $P(Y|X)$ based on samples from $P(X)$. In Figures 2(a) and 2(b) the y-coordinates of the red points correspond to $\mathbf{y}^*$ and the x-coordinates to one MCMC sample from $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$ (Eq. (3)). Given the sample $(\mathbf{x}^i, \boldsymbol{\theta}^i)$, $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$, plotted in blue, is computed by supervised inverse GP regression. In Fig. 2(a) the grey points were generated according to $Y = 2X^3 + X + E$, with X having a uniform distribution and E zero-mean Gaussian noise. On the other hand, the distribution of X in Fig. 2(b) is sub-Gaussian and the noise is not additive. In this case we often still get "good" MCMC samples.

On the contrary, in Fig. 2(c) the x-coordinates of the red points correspond to $\mathbf{x}^*$ and the y-coordinates to

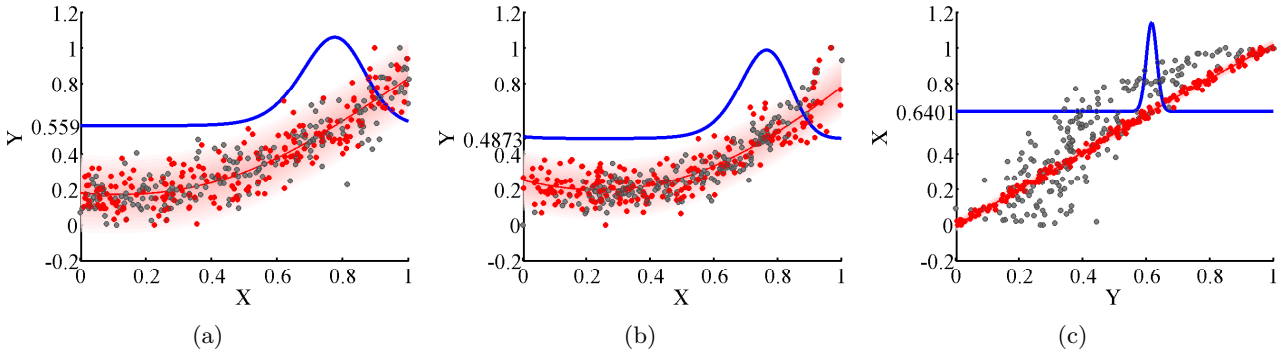

Figure 2: The grey points are generated according to $X \to Y$. (a), (c): uniform $P(X)$, additive Gaussian noise, (b): sub-Gaussian $P(X)$, non-additive noise. (a), (b): the y-coordinates of the red points correspond to $\mathbf{y}^*$ and the x-coordinates to one MCMC sample from $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$. Given the sample $(\mathbf{x}^i, \boldsymbol{\theta}^i)$, $p(x|y, \mathbf{y}^*, \mathbf{x}^i, \boldsymbol{\theta}^i)$, plotted in blue, is computed by supervised inverse GP regression. (c): note that x and y axes are exchanged. The x-coordinates of the red points correspond to $\mathbf{x}^*$ and the y-coordinates to one sample from $p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x}^*)$. Given the sample $(\mathbf{y}^i, \boldsymbol{\theta}^i)$, $p(y|X = 0.64, \mathbf{x}^*, \mathbf{y}^i, \boldsymbol{\theta}^i)$, plotted in blue, is computed by inverse regression.

one MCMC sample from $p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x}^*)$. In this case we often get "bad" MCMC samples as expected since we should not be able to estimate $P(Y|X)$ based on samples from $P(X)$ (Postulate 1).

The step of sampling from the high dimensional distribution $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$ is not trivial. Additionally, there are two modes with equal probabilities, namely, one that corresponds to the ground truth $\mathbf{x}^*$ and one to the "mirror" of $\mathbf{x}^*$ (flipping X left to right). Good initialization is crucial for sampling from this high-dimensional space. The good news is that, for the purpose of causal inference, we have the luxury of initializing the sampling algorithm with the ground truth $\mathbf{x}^*$, since this is given (but we treat it as a latent variable), and with $\boldsymbol{\theta}^*$, which is computed by maximizing the likelihood $p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$. This is fair as long as it is done for both causal directions to be checked. With this initialization, slice sampling starts from the correct mode of $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}^*)$ and usually (apart from very noisy cases), we don't get samples from the "mirror" mode. In any case, for every sample, $\mathbf{x}^*$ is used to decide to keep either this or its mirror. Initializing slice sampling with $\mathbf{x}^*$, we still get an asymmetry between cause and effect: even by initializing with the ground truth $\mathbf{x}^*$, if $Y \to X$ and we try to predict $P(X|Y)$ from $P(Y)$ (which are independent), then we eventually often get "bad" MCMC samples similar to the one in Fig. 2(c). Of course, this slice sampling initialization is only feasible for the purpose of causal inference, where both $\mathbf{x}^*$ and $\mathbf{y}^*$ are given. If the goal is just estimating $P(X|Y)$ based on samples from $P(Y)$, then we only get to see $\mathbf{y}^*$ and such a sampling initialization is not possible. In that sense, to be precise, the conditional $P(X|Y)$ is not estimated based only on $\mathbf{y}^*$, but also using some side information for $\mathbf{x}^*$ (for sampling initialization).

One final point of discussion is the choice of the hyperparameters' prior. Non-invertible functional relationships between the observed variables can provide clues to the generating causal model [Friedman and Nachman, 2000]. In contrast, in the invertible case it gets more difficult to infer the causal direction. This is one more reason to restrict $\boldsymbol{\theta}$ to favor more regular functions (of large length-scale).

6 EXPERIMENTS

6.1 Simulated Data

We generate data both with additive noise, according to $Y = f(X) + E$, with $f(X) = bX^3 + X$, and non-additive noise. Non-additive noise is simulated according to $Y = f(X) + E$, with $P(E) = \sigma\mathcal{N}(0,1)\,|\sin(2\pi\nu X)| + \frac{1}{4}\sigma\mathcal{N}(0,1)\,|\sin(2\pi(10\nu)X)|$[1]. By multiplying with a sinusoidal function the width of the noise varies for different values of X. ν controls the frequency of the wave. The results are included in Fig. 4, for a non-linear f (setting $b = 2$), and in Fig. 5, for a linear f (setting $b = 0$). The three first columns of the figures refer to data generated with additive noise and the fourth column with non-additive noise. We use four distributions for $P(X)$: standard uniform, sub-Gaussian, Gaussian and super-Gaussian, each one corresponding to one row of Figures 4 and

[1]Note that we call $Y = f(X) + E$ an additive noise model only if $X \perp\!\!\!\perp E$. This comes from the perspective of structural equations where the noise term is usually meant to be independent of X. Then a conditional $P(Y|X)$ generated by *dependent additive* noise can only be generated by a structural equation with non-additive noise.

5. For sub and super-Gaussian, data were generated from a Gaussian distribution and their absolute values were raised to the power q while keeping the original sign. $q = 0.7$ for the sub-Gaussian distribution (which is also close to bimodal), while $q = 1.3$ for the super-Gaussian. Similarly, three distributions are used for $P(E)$: sub-Gaussian, Gaussian, and super-Gaussian, each one corresponding to one of the first three columns of Figures 4 and 5. The x-axis of the first three columns refers to the standard deviation (std) of the noise. Three values of std are used: $0.25, 0.45$ and 0.8, each multiplied by the standard deviation of $f(X)$, in order to get comparable results across different experiments. The x-axis of the fourth column is the frequency of the sinusoidal wave, ν, with values from $\{4, 0.5, 025\}$. We generate $N = 200$ samples for each simulated setting.

We compare the proposed causal inference method (CURE) with some of the causal inference methods reviewed in the introduction: additive noise models (ANM) [Hoyer et al., 2009, Peters et al., 2014], information-geometric causal inference (IGCI) [Daniusis et al., 2010, Janzing et al., 2012] and Bayesian model selection (GPI) [Mooij et al., 2010]. CURE uses a uniform prior so a preprocessing step is first applied to X and Y to remove possible isolated points (low-density points). For CURE, $M = 15000$ MCMC samples are generated from the 203-dimensional ($N = 200$) posterior using the slice sampling method, from which the first 5000 are discarded. Since it is difficult to sample from this very high-dimensional space, to get a more robust answer, we report the average $D_{X|Y}$ and $D_{Y|X}$ across 4 repetitions of CURE for each dataset. We call those repetitions "internal" repetitions of the CURE algorithm to distinguish them from the repetitions of the simulations. Assume $D^i_{X|Y}$ is the output of the i^{th} internal repetition. Then, $D_{X|Y} = \frac{1}{4}\sum_{i=1}^{4} D^i_{X|Y}$ and $D_{Y|X} = \frac{1}{4}\sum_{i=1}^{4} D^i_{Y|X}$. We conduct 20 repetitions for each combination of method and simulation setting, apart from CURE which is repeated 10 times, due to the high computational complexity of the MCMC sampling step. The y-axis of Figures 4 and 5 corresponds to the percentage of correct causal inferences.

For non-linear f (Fig. 4), we can observe that CURE (red) infers correctly the causal direction when $P(X)$ is uniform or sub-Gaussian and for all noise distributions. The accuracy degrades in some cases of Gaussian and super-Gaussian $P(X)$ (due to the uniform prior) with high standard deviation of $P(E)$. IGCI (green) infers the causal direction correctly in almost all cases, even though it was proposed for deterministic relations. ANM (blue) gets 100% correct decisions on the additive noise data, however, its performance is really degraded when it comes to non-additive noise. Finally, GPI (brown) performs better with uniform $P(X)$ than with Gaussian or super-Gaussian, where its results are worse compared to the other methods.

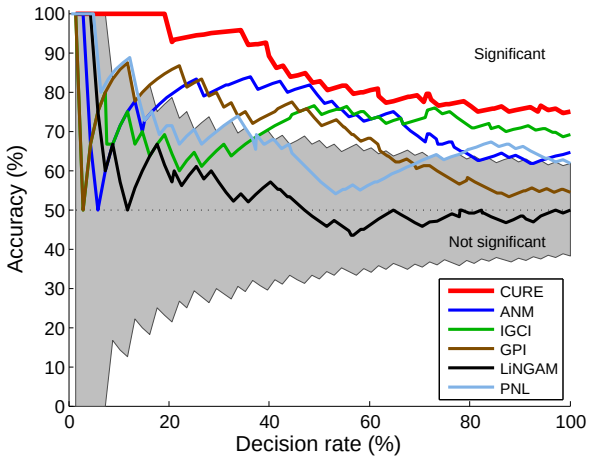

Figure 3: Results of various causal inference methods for 81 cause-effect pairs (86 excluding 5 multivariate pairs), showing the percentage of correct causal inferences for each decision rate.

For the linear case (Fig. 5), the performance of almost all methods gets worse since it gets more difficult to recover the causal direction. Specifically, the case of linear f and Gaussian $P(X)$ and $P(N)$ is non-identifiable [Hoyer et al., 2009]. This is also supported by the results: in this case the decision of all methods is close to 50% (random guess). For uniform $P(X)$, CURE outperforms the other methods, however for non-uniform $P(X)$ its performance often degrades. ANM generally performs relatively well with additive noise, however, it again fails in the non-additive noise case. GPI performs much better in the linear compared to the non-linear case, outperforming the other methods in several cases. Finally, IGCI often fails in the linear case.

6.2 Real Data

Further, we evaluate the performance of our method on real-world data, namely on a database with cause-effect pairs[2] (version 0.9), a detailed description of which was recently provided by Mooij et al. [2014]. It consists of 86 pairs of variables from various domains with known causal structure, the first 41 of which are from the UCI Machine Learning Repository [Bache and Lichman, 2013]. The task is to infer the causal direction for each of the pairs. Each

[2] http://webdav.tuebingen.mpg.de/cause-effect/

pair is weighted as suggested in the database. Five of the pairs have multivariate X or Y and were excluded from the analysis. At most $N = 200$ samples from each cause-effect pair are used (less than 200 only if the pair itself has less samples). For CURE, $M = 10000$ MCMC samples are generated, after burning the first 10000 samples and additionally discarding every other sample. The average $D_{X|Y}$ and $D_{Y|X}$ across 8 internal repetitions of CURE are computed for each dataset. Two more methods participate in this comparison: Post-Nonlinear Models (PNL) [Zhang and Hyvärinen, 2009] and Linear Non-Gaussian Acyclic Models (LiNGAM) [Shimizu et al., 2006]. The results for all the methods are depicted in Fig. 3. The y-axis corresponds to the percentage of correct causal inferences. As the causal inference methods we compare with, we also output a ranking of the pairs according to a confidence criterion along with the decisions on the causal direction. The method is more certain about the decided direction of the top-ranked pairs as opposed to the low-ranked ones. Using this ranking, we can decide on the causal direction of only a subset of the pairs for which we are more confident about. This way, we trade off accuracy versus the number of decisions taken. The x-axis of Fig. 3 corresponds to the percentage of pairs for which we infer the causal direction (100% means that we are forced to decide upon the direction of all 81 pairs). A good confidence criterion corresponds to the accuracy being lowest for decision rate 100% and increase monotonically as the decision rate decreases. As a confidence criterion we choose to use the ratio between $\sigma_{D_{X|Y}} = std(\{D^i_{X|Y}\}_{1\leq i\leq 8})$ and $\sigma_{D_{Y|X}} = std(\{D^i_{Y|X}\}_{1\leq i\leq 8})$, with denominator the one that corresponds to the inferred causal direction (smaller D). The idea is that, if $X \to Y$ and we try to predict $P(X|Y)$ based on $P(Y)$, the empirical variance of the algorithm across internal repetitions is expected to be small: MCMC samples are expected to correspond to conditionals close to the ones of the ground truth. On the other hand, when predicting $P(Y|X)$ based on $P(X)$ (which are independent), the variance is higher across internal repetitions.

We consider the null hypothesis that "the causal inference algorithm outputs random decisions with probability 1/2 each". Then the grey area of Fig. 3 indicates the 95% confidence interval of a binomial distribution with n trials where n is the (weighted) number of cause-effect pairs (the weights given as suggested in the database). Thus, the area outside the grey area corresponds to results significantly correlated with the ground truth. We can observe that CURE (bold red) outperforms the other methods for all decision rates, however it is difficult to draw any definite conclusions about the relative performance of these methods based on only 81 cause-effect pairs. Moreover, the ratio of standard deviations that is used as a confidence criterion for CURE seems to be a good choice: for low decision rates we even get 100% accuracy, decreasing more or less monotonically as the decision rate increases. IGCI performs well for high decision rates but its confidence criterion does not behave as expected. ANM has a better confidence criterion, however, its performance is quite low compared to CURE and IGCI when it is forced to take a decision. The result of PNL is marginally significant in the forced-decision regime. Finally, the results of GPI and LINGAM are not significantly correlated with the ground truth in the forced-decision regime.

Increasing N, the performance is obviously increasing. For example, running ANM with all the available samples of the 81 cause-effect pairs results in an accuracy of 72% [Peters et al., 2014], much higher than its accuracy with $N = 200$ (Fig. 3). Unfortunately, the computational complexity of CURE did not allow for it to be run for such a big sample size (thousands for some pairs). However, we consider very encouraging the fact that CURE can yield accuracy 75% already with $N = 200$.

7 CONCLUSION

We proposed a method (CURE) to infer the causal direction between two random variables given a sample from their joint distribution. It is based on the postulate that the marginal distribution of the cause and the conditional distribution of the effect given the cause contain no information about each other. In contrast, the distribution of the effect and the conditional of the cause given the effect may share information. Exploiting this asymmetry, if we can estimate $P(X|Y)$ based on $P(Y)$ more accurately than $P(Y|X)$ based on $P(X)$, then $X \to Y$, is inferred. Otherwise, $Y \to X$ is inferred. For that, unsupervised inverse GP regression was proposed as a method to estimate a conditional from samples from the corresponding marginal. CURE was evaluated in both simulated and real data, and found to perform well compared to existing methods. In particular, it outperforms five existing causal inference methods on our real data experiments. A downside is the comparably high computational cost due to the large number of required MCMC steps.

Acknowledgements

We would like to thank Kun Zhang and Joris Mooij for helpful discussions.

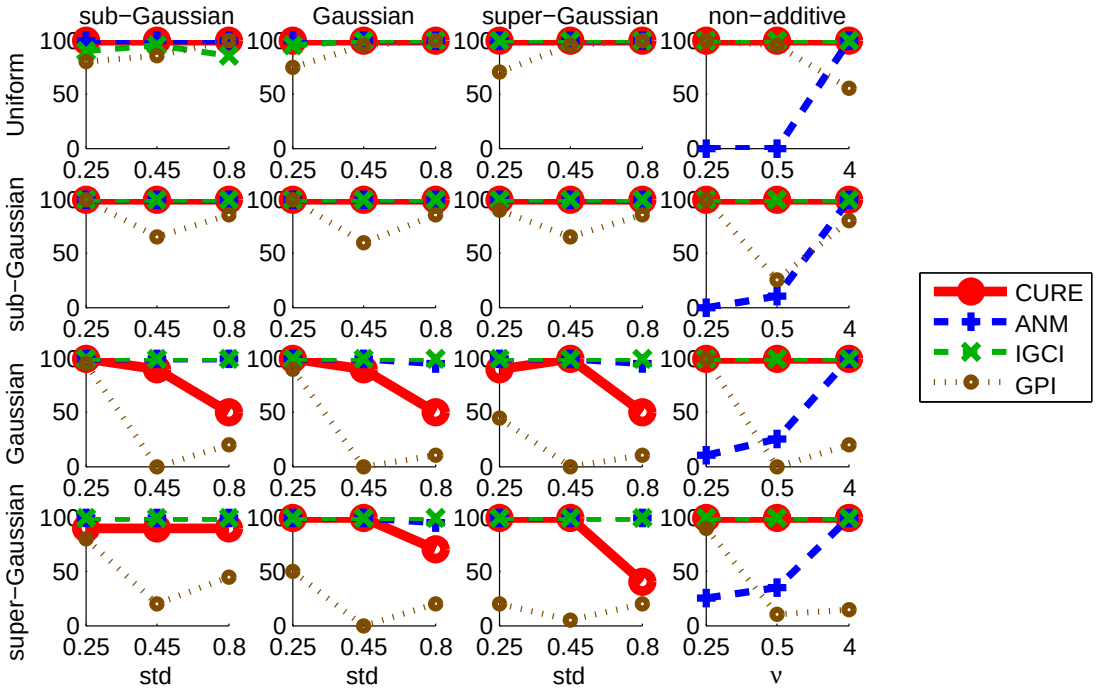

Figure 4: Performance (percentage of correct causal inferences) of various causal inference methods for simulated data with a non-linear function f. Rows correspond to the distribution of the cause, $P(X)$. The three first columns correspond to the distribution, $P(E)$, of the additive noise term, with the x-axis referring to 3 different standard deviations of the noise. The last column corresponds to non-additive noise, with the x-axis referring to 3 different frequencies of the sinusoidal wave (used to simulate non-additive noise).

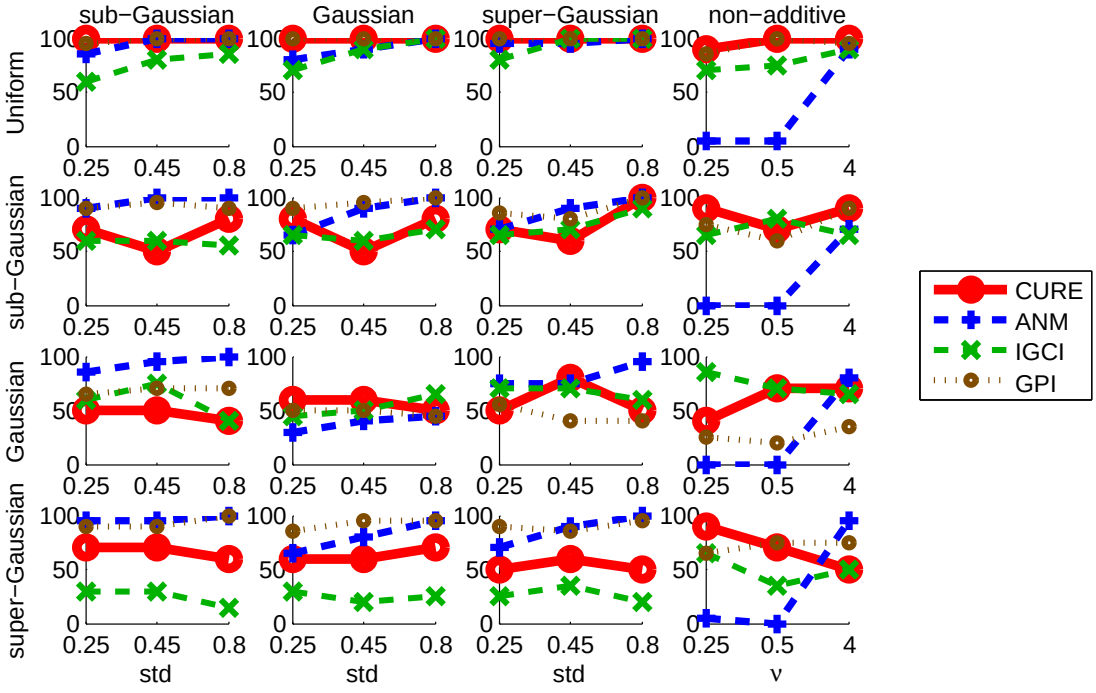

Figure 5: As in Fig. 4 but with a linear function f.

References

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

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 21 (NIPS)*, 2009.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems 23 (NIPS)*, 2010.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2010.

D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

C. E. Rasmussen and Williams C. K. I. *Gaussian processes for machine learning*. MIT Press, 2006.

M. Titsias and N. Lawrence. Bayesian gaussian process latent variable model. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

R. M. Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

N. Friedman and I. Nachman. Gaussian process networks. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv:1412.3773*, 2014.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.