On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model

KUN ZHANG, Max-Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
ZHIKUN WANG, Max-Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
JIJI ZHANG, Department of Philosophy, Lingnan University, Hong Kong
BERNHARD SCHÖKOPF, Max-Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

Compared to constraint-based causal discovery, causal discovery based on functional causal models is able to identify the whole causal model under appropriate assumptions [Shimizu et al. 2006; Hoyer et al. 2009; Zhang and Hyvärinen 2009b]. Functional causal models represent the effect as a function of the direct causes together with an independent noise term. Examples include the linear non-Gaussian acyclic model (LiNGAM), nonlinear additive noise model, and post-nonlinear (PNL) model. Currently there are two ways to estimate the parameters in the models; one is by dependence minimization, and the other is maximum likelihood. In this paper, we show that for any acyclic functional causal model, minimizing the mutual information between the hypothetical cause and the noise term is equivalent to maximizing the data likelihood with a flexible model for the distribution of the noise term. We then focus on estimation of the PNL causal model, and propose to estimate it with the warped Gaussian process with the noise modeled by the mixture of Gaussians. As a Bayesian nonparametric approach, it outperforms the previous one based on mutual information minimization with nonlinear functions represented by multilayer perceptrons; we also show that unlike the ordinary regression, estimation results of the PNL causal model are sensitive to the assumption on the noise distribution. Experimental results on both synthetic and real data support our theoretical claims.

Categories and Subject Descriptors: I.2.0 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms: Systems and Information Theory, Learning, Probability and Statistics

Additional Key Words and Phrases: Causal discovery, functional causal model, post-nonlinear causal model, statistical independence, maximum likelihood

1. INTRODUCTION

There has been a long history of debate on causality in philosophy, statistics, machine learning, data mining, and related fields. In particular, people have been concerned with the causal discovery problem, i.e., how to discover causal information from purely observational data. Traditionally, it has been noted that under the causal Markov condition and the faithfulness assumption, based on conditional independence relationships of the variables, one could recover an equivalence class of the underlying causal structure [Spirtes et al. 2001; Pearl 2000]. This approach involves conditional independence tests [Zhang et al. 2011], which would be a difficult task if the form of dependence is unknown. Furthermore, the solution of this approach for causal discovery is usually non-unique, and in particular, it does not help in the two-variable case, where no conditional independence relationship is available.

Recently several causal discovery approaches based on functional causal models have been proposed. A functional causal model represents the effect Y as a function of the direct causes X and some unmeasurable noise:

$$Y = f(X, N; \boldsymbol{\theta}_1), \tag{1}$$

where N is the noise that is assumed to be independent from X, the function $f \in \mathcal{F}$ explains how Y is generated from X, $\mathcal{F}$ is an appropriately constrained functional class, and $\boldsymbol{\theta}_1$ is the parameter set involved in f. We assume that the transformation from (X, N) to (X, Y) is invertible, such that N can be uniquely recovered from the observed variables X and Y.

For convenience of presentation, let us assume that both X and Y are one-dimensional variables. Without precise knowledge on the data-generating process, the functional causal model should be flexible enough such that it could be adapted to approximate the true data-generating process; more importantly, the causal direction implied by the functional causal model has to be identifiable, i.e., the model assumption, especially the independence between the noise and cause, holds for only one direction, such that it implies the causal asymmetry between X and Y. Under the above conditions, one can then use functional causal models to determine the causal direction between two variables, given that they have a direct causal relationship in between and do not have any confounder: for both directions, we fit the functional causal model, and then test for independence between the estimated noise and the hypothetical cause, and the direction which gives independent noise is considered plausible.

Several functional causal models have been shown to be able to produce unique causal directions, and have received practical applications. In the linear, non-Gaussian, and acyclic model (LiNGAM [Shimizu et al. 2006]), f is linear, and at most one of the noise N and cause X is Gaussian. The nonlinear additive noise model [Hoyer et al. 2009; Zhang and Hyvärinen 2009a] assumes that f is nonlinear with additive noise N. In the post-nonlinear (PNL) causal model [Zhang and Hyvärinen 2009b], the effect Y is further generated by a post-nonlinear transformation on the nonlinear effect of the cause X plus noise N:

$$Y = f_2(f_1(X) + N), \tag{2}$$

where both f_1 and f_2 are nonlinear functions and f_2 is assumed to be invertible. As in post-nonlinear independent component analysis [Taleb and Jutten 1999; Zhang and Chan 2005], the post-nonlinear transformation f_2 represents the sensor or measurement distortion, which is frequently encountered in practice. In particular, the PNL causal model has a very general form (the former two are its special cases), but it has been shown to be identifiable in the general case (except five specific situations given in [Zhang and Hyvärinen 2009b]).

In this paper we are concerned with distinguishing cause from effect of two continuous variables based on functional causal models. Developing practical methods for causal discovery of more than two variables is an important step towards solving large-scale real-world causal analysis problems, but since we are interested in certain fundamental theoretical issues raised in estimating functional causal models, we limit ourselves to the two-variable case for the sake of clarity.[1] One should also pay attention to such theoretical issues when developing causal discovery methods for more than two variables. Causal discovery of discrete variables or of mixed discrete and continuous variables might require different classes of functional causal models, and are not discussed here.

We aim to clarify several crucial issues in estimating the functional causal models, and discuss the practical implications of such theoretical studies. For causal discovery based on the nonlinear additive noise model, some regression methods have been proposed to directly minimize the dependence between noise and the hypothetical cause [Mooij et al. 2009; Yamada and Sugiyama 2010]. Such methods only apply to the additive noise model, and model selection is usually not well-founded. As the first contribution, here we show that for any functional causal model, in which noise

[1] One way to estimate the causal model on more than two variables based on functional causal models is to use exhaustive search: for all possible causal orderings, fit functional causal models for all hypothetical effects separately, and then do model checking by testing for independence between the estimated noise and the corresponding hypothetical causes. However, note that the complexity of this procedure increases super-exponentially along with the number of variables, and hence smart approaches are needed. This is beyond the scope of this paper.

is not necessarily additive, minimizing the mutual information between noise and the predictor is equivalent to maximizing the data likelihood, given that the noise model is flexible.

For estimating some statistical models, such as linear regression, the Gaussianity assumption on the data gives statistically consistent estimators; moreover, the Gaussianity assumption usually makes the estimation procedure computationally efficient. However, there are other statistical models for which one has to consider the true data distribution to derive statistically consistent estimators. For instance, when estimating the independent component analysis (ICA [Hyvärinen et al. 2001]) model, one has to make use of the non-Gaussianity of the data; otherwise the model is not identifiable. We are interested in whether simply using a Gaussian distribution for the noise gives a consistent estimator of the functional causal model. If this is not the case, using the Gaussian distribution might give misleading results. Thus, as the second contribution, we show that for estimation of the functional causal model where noise is not additive, the solution depends on the assumption on the noise distribution.

These results motivate the use of Bayesian inference to estimate the functional causal model with a flexible noise model. As noted above, the PNL causal model has a very general form, and yet it allows the causal direction to be identifiable in the general case. Finally, to give some practical implications of the above theoretical results, we focus on this causal model; we propose to estimate it by warped Gaussian processes with the noise distribution represented by the mixture of Gaussians (MoG), and compare it against warped Gaussian processes with the Gaussian noise and mutual information minimization approach with nonlinear functions represented by multi-layer perceptrons (MLPs) [Zhang and Hyvärinen 2009b]. The empirical results illustrate the necessity of adopting a flexible noise model, instead of a Gaussian one, and further demonstrate that the maximum likelihood framework, compared to the mutual information minimization one, might provide a more natural way to learn hyperparameters in the model.[2]

2. ASYMMETRY OF CAUSE AND EFFECT IN FUNCTIONAL CAUSAL MODELS

In this section we explain why f in the functional causal model (1) has to be properly constrained, and then give some examples of the functional forms f, including the PNL causal model.

2.1. General Claims

Given any two random variables X and Y with continuous support, one can always construct another variable, denoted by $\tilde{N}$, which is statically independent from X, as suggested by the following lemma.

LEMMA 1. *For any two variables X and Y with continuous support, let $F_{Y|X}$ be the conditional cumulative distribution function of Y given X and q be an arbitrary continuous and strictly monotonic function with a non-zero derivative. The quantity $\tilde{N} = q \circ F_{Y|X}$, where $\circ$ denotes function composition, is then always independent from X. Furthermore, the transformation from $(X, Y)^T$ to $(X, \tilde{N})^T$ is always invertible, in the sense that Y can be uniquely reconstructed from $(X, \tilde{N})^T$.*

PROOF. Consider $F_{Y|X}$ as a random variable. Since at any possible value of X, $F_{Y|X}$ is always uniformly distributed, we know that $F_{Y|X}$ is statistically independent from X; $\tilde{N} = q \circ F_{Y|X}$ is then also independent from X. (One may refer to [Hyvärinen and

[2]A preliminary and shorter version of this paper was presented at The First IEEE/ICDM Workshop on Causal Discovery [Zhang et al. 2013b].

Pajunen 1999] for the detailed procedure to construct $\tilde{N}$.) Moreover, the invertibility can be seen from the fact that the determinant of the transformation from $(X, Y)^T$ to $(X, \tilde{N})^T$, which is $q' \cdot p(Y|X)$,[3] is positive everywhere on the support, under the conditions specified in Lemma 1. □

Let $\tilde{N}$ be the noise term N in the functional causal model (1), and one can see that without constraints on f, there always exists the function f such that the independence condition on N and X holds. Similarly, we can always represent X as a function of Y and an independent noise term. That is, any two variables would be symmetric according to the functional causal model, if f is not constrained. Therefore, in order for the functional causal models to be useful to determine the causal direction, we have to introduce certain constraints on the function f such that the independence condition on noise and the hypothetical cause holds for only one direction.

2.2. Examples

For simplicity let us assume that the true causal direction is $X \to Y$. The functional class $\mathcal{F}$ is expected to be able to approximate the data generating process, but very importantly, it should be well constrained such that noise cannot be independent from the assumed cause for the backward direction. A simple choice for $\mathcal{F}$ is a linear model, i.e., $Y = \mu + \alpha X + N$, where μ is a constant. It has been shown that under the condition that in the data generating process at most one of N and X is Gaussian, Y and N_Y in the backward direction are always dependent [Shimizu et al. 2006]; this motivated the so-called linear, non-Gaussian, and acyclic model (LiNGAM).

In practice nonlinearity is rather ubiquitous in the data generating process, and should be taken into account in the functional class. A very general setting for $\mathcal{F}$ is given by the PNL causal model [Zhang and Hyvärinen 2009b]; see (2). It has been shown that for the PNL causal model, except in several special cases (including the linear-Gaussian case discussed above), in the backward direction N_Y is always dependent on Y, so that one can find the plausible causal direction with an independent noise term. If f_2 in the PNL causal model is the identity mapping, this model reduces to the additive noise model [Hoyer et al. 2009].

3. RELATIONSHIP BETWEEN DEPENDENCE MINIMIZATION AND MAXIMUM LIKELIHOOD

Let us now suppose that both X and Y are continuous and that X is the direct cause of Y; we have assumed that both X and Y are one-dimensional and that there is no common cause for X and Y. The main result will also apply when X contains multiple variables, as we shall see later.

We consider the functional causal model (1). Denote by $p(X, Y)$ the true density of (X, Y), and by $p_{\mathcal{F}}(X, Y)$ the joint density implied by (1). The model (1) assumes $p(X, N) = p(X)p(N)$; because the Jacobian matrix of the transformation from $(X, N)^T$ to $(X, Y)^T$ is

$$\mathbf{J}_{X \to Y} = \begin{pmatrix} \frac{\partial X}{\partial X} & \frac{\partial X}{\partial N} \\ \frac{\partial Y}{\partial X} & \frac{\partial Y}{\partial N} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{\partial f}{\partial X} & \frac{\partial f}{\partial N} \end{pmatrix}, \tag{3}$$

the absolute value of its determinant is $|\mathbf{J}_{X \to Y}| = |\frac{\partial f}{\partial N}|$, and hence we have

$$P_{\mathcal{F}}(X, Y) = p(X, N)/|\mathbf{J}_{X \to Y}| = p(X)p(N)\left|\frac{\partial f}{\partial N}\right|^{-1}, \tag{4}$$

which implies $P_{\mathcal{F}}(Y|X) = P_{\mathcal{F}}(X, Y)/p(X) = p(N)\left|\frac{\partial f}{\partial N}\right|^{-1}$.

[3]For notational convenience, we write $p_{Y|X}(y|x)$ as $p(Y|X)$.

Now let us introduce the concept of mutual information [Cover and Thomas 1991]. As a canonical measure of statistical dependence, mutual information between X and N is defined as:

$$\begin{aligned} I(X,N) &= \int p(X,N)\log\frac{p(X,N)}{p(X)p(N)}dxdn \\ &= -\mathbb{E}\log p(X) - \mathbb{E}\log p(N) + \mathbb{E}\log p(X,N), \end{aligned} \tag{5}$$

where $\mathbb{E}(\cdot)$ denotes expectation. $I(X,N)$ is always non-negative and is zero if and only if X and N are independent.

When X contains multiple variables, $\frac{\partial X}{\partial X}$ in (3) becomes the identity matrix; in this case, $p(X)$ is the joint distribution of all components of X, and (4) as well as (5) still holds.

3.1. Maximum likelihood and dependence minimization for functional causal models

Suppose we fit the model (1) on the given sample $\mathcal{D} \triangleq \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^T$; as the transformation from (X,N) to (X,Y) is invertible, given any parameter set $\boldsymbol{\theta}_1$ involved in the function f, the noise N can be recovered, and we denote by $\hat{N}$ the estimate. We further denote by $\boldsymbol{\theta}_2$ the parameter set in $p(N)$. We are now ready to show that the attained likelihood of (1) is directly related to the dependence between the estimated noise N and X.

For any parameter set $\boldsymbol{\theta} \triangleq (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, the log-likelihood attained by the model (1) is

$$\begin{aligned} l_{X\to Y}(\boldsymbol{\theta}) &= \sum_{i=1}^T \log P_{\mathcal{F}}(\mathbf{x}_i, \mathbf{y}_i) \\ &= \sum_{i=1}^T \log p(X=\mathbf{x}_i) + \sum_{i=1}^T \log p(N=\hat{\mathbf{n}}_i;\boldsymbol{\theta}_2) - \sum_{i=1}^T \log\Big|\frac{\partial f}{\partial N}\Big|_{X=\hat{\mathbf{x}}_i, N=\hat{\mathbf{n}}_i}\Big|. \end{aligned} \tag{6}$$

On the other hand, the mutual information (sample version) between X and $\hat{N}$ for the given parameter set $\boldsymbol{\theta}$ is

$$\begin{aligned} &I(X,\hat{N};\boldsymbol{\theta}) \\ &= -\frac{1}{T}\sum_{i=1}^T \log p(X=\mathbf{x}_i) - \frac{1}{T}\sum_{i=1}^T \log p(\hat{N}=\hat{\mathbf{n}}_i;\boldsymbol{\theta}_2) + \frac{1}{T}\sum_{i=1}^T \log\Big|\frac{\partial f}{\partial N}\Big|_{X=\hat{\mathbf{x}}_i, N=\hat{\mathbf{n}}_i}\Big| + \\ &\qquad\qquad \frac{1}{T}\sum_{i=1}^T \log p(X=\mathbf{x}_i, Y=\mathbf{y}_i), \end{aligned} \tag{7}$$

where the last term does not depend on $\boldsymbol{\theta}$ and can then be considered as constant. We then have the following result on the relationship between $l_{X\to Y}(\boldsymbol{\theta})$ and $I(X,\hat{N};\boldsymbol{\theta})$ defined above.

THEOREM 2. *For the model (1) with any value of the parameter set $\boldsymbol{\theta}$, $l_{X\to Y}(\boldsymbol{\theta})$, defined in (6), and $I(X,\hat{N};\boldsymbol{\theta})$, defined in (7), are related in the following way:*

$$\frac{1}{T} l_{X\to Y}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{i=1}^T \log p(X=\mathbf{x}_i, Y=\mathbf{y}_i) - I(X,\hat{N};\boldsymbol{\theta}). \tag{8}$$

Therefore, the parameter set $\boldsymbol{\theta}^*$ that maximizes the likelihood of the model (1) also minimizes the mutual information $I(X,\hat{N})$.

PROOF. (6) directly follows (4), and now we prove (7). Note that the absolute value of the determinant of the transformation from (X, Y) to $(X, \hat{N})$ is $|\mathbf{J}_{Y\to X}| = |\mathbf{J}_{X\to Y}|^{-1}$. Recalling $|\mathbf{J}_{X\to Y}| = \left|\frac{\partial f}{\partial N}\right|$, consequently, we have $p(X, \hat{N}) = p(X, Y)/|\mathbf{J}_{Y\to X}| = p(X, Y)\left|\frac{\partial f}{\partial N}\right|$.

According to (5), one can see

$$I(X, \hat{N}; \boldsymbol{\theta}) = -\mathbb{E}\log p(X) - \mathbb{E}\log p(\hat{N}) + \mathbb{E}\Big\{\log p(X, Y) + \log\Big|\frac{\partial f}{\partial N}\Big|\Big\},$$

whose sample version is (7). (8) can be directly seen from (6) and (7). □

Theorem 2 is a consequence of (4) and (5), which also hold when X contains multiple variables. Therefore, the above result also applies when X is a random vector.

We then consider the likelihood of the direction $Y \to X$ can attain, denoted by $l_{Y\to X}$. That is, we fit the sample with the model

$$X = g(Y, N_Y; \boldsymbol{\psi}) \tag{9}$$

where $g \in \mathcal{F}$, N_Y is assumed to be independent from Y, and $\boldsymbol{\psi}$ is the parameter set. We shall show that *if the functional class $\mathcal{F}$ is appropriately chosen such that X is independent from N (i.e., (1) holds), but the reverse model (9) does not hold, i.e., these does not exist $g \in \mathcal{F}$ such that N_Y is independent from Y in (9), one can then determine the causal direction with the likelihood principle*. In fact, the maximum likelihood attained by the former model is higher than that of the latter, as seen from the following theorem.

THEOREM 3. *Let $\boldsymbol{\theta}^*$ denote the maximum likelihood estimator of the parameters in (1) on the given sample $\mathcal{D}$. Similarly, let $\boldsymbol{\psi}^*$ be the maximum likelihood estimator of the parameters in the model (9) on $\mathcal{D}$. Assume that the model (1) is true, in the sense that N is independent from X. Further assume that the model (9) does not hold, in the sense that when estimating (9) with maximum likelihood, as $T \to \infty$, the resulting noise $\hat{N}_Y$ is dependent on Y, i.e., $I(Y, \hat{N}_Y; \boldsymbol{\psi}^*) > 0$ as $T \to \infty$.*

Then as $T \to \infty$, the maximum likelihood $l_{X\to Y}(\boldsymbol{\theta}^*)$ is higher than $l_{Y\to X}(\boldsymbol{\psi}^*)$, and the difference is

$$\frac{1}{T}l_{X\to Y}(\boldsymbol{\theta}^*) - \frac{1}{T}l_{Y\to X}(\boldsymbol{\psi}^*) = I(Y, \hat{N}_Y; \boldsymbol{\psi}^*). \tag{10}$$

PROOF. According to (8), we have

$$\frac{1}{T}l_{X\to Y}(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{i=1}^{T}\log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(X, \hat{N}; \boldsymbol{\theta}^*), \tag{11}$$

$$\frac{1}{T}l_{Y\to X}(\boldsymbol{\psi}^*) = \frac{1}{T}\sum_{i=1}^{T}\log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(Y, \hat{N}_Y; \boldsymbol{\psi}^*). \tag{12}$$

Bearing in mind that $I(X, \hat{N}; \boldsymbol{\theta}^*) \to 0$ as $T \to \infty$, one substracts (12) from (11) and obtains (10). □

3.2. Loss Caused by a Wrongly Specified Noise Distribution

As claimed in Theorem 2, estimating the functional causal model (1) by maximum likelihood or mutual information minimization aims to maximize

$$J_{X\to Y} = \sum_{i=1}^{T} \log p(N = \hat{\mathbf{n}}_i) - \sum_{i=1}^{T} \log \Big| \frac{\partial f}{\partial N}\big|_{X=\hat{\mathbf{x}}_i, N=\hat{\mathbf{n}}_i} \Big|. \tag{13}$$

3.2.1. Theoretical results. In the linear model (i.e., f in (1) is a linear function of X plus the noise term N) or the nonlinear additive noise model (i.e., f is a nonlinear function of X plus N), $\frac{\partial f}{\partial N} \equiv 1$, and the above objective function reduces to $\sum_{i=1}^{T} \log p(N = \hat{\mathbf{n}}_i)$, whose maximization further reduces to the ordinary regression problem. It is well known that in such situations, if N is non-Gaussian, parameter estimation under the Gaussianity assumption on N is still statistically consistent.

However, it is important to note that this might not be the case for the general functional causal models. In fact, Bickel and Doksum [Bickel and Doksum 1981] investigated the statistical consistency properties of the parameters in the Box-Cox transformation, which is a special case of the PNL formulation (2) where f_1 is linear and f_2 is in a certain nonlinear form. They found that if the noise distribution is wrongly specified, one cannot expect consistency of the estimated parameters in the Box-Cox transformation.

Roughly speaking, if the noise distribution is set to a wrong one, one cannot guarantee the consistency of the estimated f for the functional causal models where $\frac{\partial f}{\partial N}$ is not constant, for instance, for the PNL causal model (2), where $\frac{\partial f}{\partial N} = f_2'$ is not constant if the post-nonlinear transformation f_2 is nonlinear. Theoretical proof is very lengthy, and here we give an intuition. If $p(N)$ is wrongly specified, the estimated f is not necessarily consistent: in this situation, compared to the true solution, the estimated f might have to be distorted in order to make the estimated noise closer to the specified distribution such that the first term in (13) becomes bigger; consequently, (13), a tradeoff of the two terms, is maximized. This will be illustrated by the following example and by simulations in Section 5.

3.2.2. An Illustrative Example. Besides the above theoretical results, it is helpful to use an example to illustrate the inconsistency of the estimated f caused by misspecification of the noise distribution.

Let the true generating process from X to Y be represented by the PNL causal model given in (2) with the following simple settings:

$$f_1^*(X) = X, \quad f_2^*(Z) = Z. \tag{14}$$

In words, the data-generating process is

$$Y = f^*(X, N) = X + N, \tag{15}$$

where N is the noise term. The true distribution of N used for data generation, $p^*(N)$, will be specified later.

Suppose we fit the following parametric functions as well as the assumed noise distribution on the data:

$$f_1(X) = (k+1)X, \quad f_2(Z) = \frac{b}{1 + e^{-c(Z-a)}} - d, \text{ and } N \sim \mathcal{N}(0, 1), \tag{16}$$

where $c > 0$. That is, the assumed causal model is

$$Y = f(X, N) = \frac{b}{1 + e^{-c[(k+1)X+N-a]}} - d. \tag{17}$$

We have certain constraints on the parameters, in particular, $d > -\min_i(\mathbf{y}_i)$, $b > \max_i(\mathbf{y}_i) + d$, to ensure that all Y values on the data set are in the codomain (or the range of Y) of the above function.

Note that the above function could successfully approximate the true data generating process (15) with an appropriate choice of the involved parameters: let $c \to 0^+$, $b = 4/c$, $d = 2/c$, $k = 0$, and a satisfy $\frac{4}{c(1+e^{ca})} - \frac{2}{c} = 0$, one can see $\frac{\partial f_2}{\partial Z} \to 1$ for small Z values and that $f_2(0) = 0$; in words, when $c \to 0^+$,f_2 in (16) is approximately an identify mapping for small Z values, and f_1 is an identity mapping since $k = 0$. Equation (17) then becomes $X + N$ when their values are not very large, that is, it reduces to the true process (15). On the other hand, if c is not close to zero, f_2 in (16) is nonlinear in Z, meaning that the fitted model (17) is different from the true one (15).

Let us check if we can recover the true data-generating process (15) by fitting the causal model (17) in which $N \sim \mathcal{N}(0, 1)$. (17) implies that

$$\begin{aligned} N &= a - \frac{1}{c}\log\Big(\frac{b}{Y+d} - 1\Big) - (k+1)X, \\ \log\Big|\frac{\partial f}{\partial N}\Big| &= -c(Z-a) - 2\log(1+e^{-c(Z-a)}) + \log bc \\ &= \log\Big(\frac{b}{Y+d} - 1\Big) + 2\log(Y+d) - \log b + \log c. \end{aligned}$$

Recall that we set $N \sim \mathcal{N}(0, 1)$ in (17). To estimate the involved parameters a, b, c, d, and k, we maximize (13); equivalently, the following quantity is to be minimized:

$$\begin{aligned} \hat{J} = \frac{1}{2}\sum_{i=1}^{T}\Big[a - \frac{1}{c}\log\Big(\frac{b}{\mathbf{y}_i+d} - 1\Big) - (k+1)\mathbf{x}_i\Big]^2 + \\ \sum_{i=1}^{T}\Big[\log\Big(\frac{b}{\mathbf{y}_i+d} - 1\Big) + 2\log(\mathbf{y}_i+d) - \log b + \log c\Big]. \end{aligned}$$

We note that it is not easy to find close-form solutions to the parameters. We therefore use numerical solutions obtained by the MATLAB 'fmincon' toolbox. Since we aim to illustrate that the parameter estimate might be inconsistent when $p(N)$ is wrongly specified, we use a large sample size, $T = 10^5$, and various settings for $p^*(N)$. Specifically, we let

$$p^*(N) = \alpha\mathbf{Exp}^c(1) + (1-\alpha)\mathcal{N}(0,1), \tag{18}$$

where $\mathrm{Exp}^c(1)$ denotes the centered exponential distribution with $\lambda = 1$, and $0 \le \alpha \le 1$. For arbitrary α, $p^*(N)$ always has a zero mean and unit variance. That is, $p^*(N)$ and $p(N)$ used in (17) have the same mean and variance, and they differ in the shape when $\alpha > 0$. We vary α from 0 to 1, and Fig. 1 shows the estimated values of c and k. Recall that if the true process is consistently estimated, $\hat{c}$ should be be very small, $\hat{k}$ is expected to be 0. In this figure, one can see that when α becomes larger, $\hat{c}$ becomes larger, and $\hat{k}$ tends to be further away from 0. Roughly speaking, the more $\hat{c}$ and $\hat{k}$ deviate from 0, the more different the estimated functional causal model is from the true one.

Fig. 2 shows the estimation results when the noise distribution is truly Gaussian (α=0, see top panels of the figure) and when it is exponentially distributed ($\alpha = 1$, bottom panels). Not surprisingly, when the noise is Gaussian, both f_1 and f_2 are accurately estimated. However, when the true noise distribution is an Exponential one, one can see that f_2 is no longer a linear function; as a consequence, the estimated noise is closer to Gaussian, as seen from Fig. 2f.

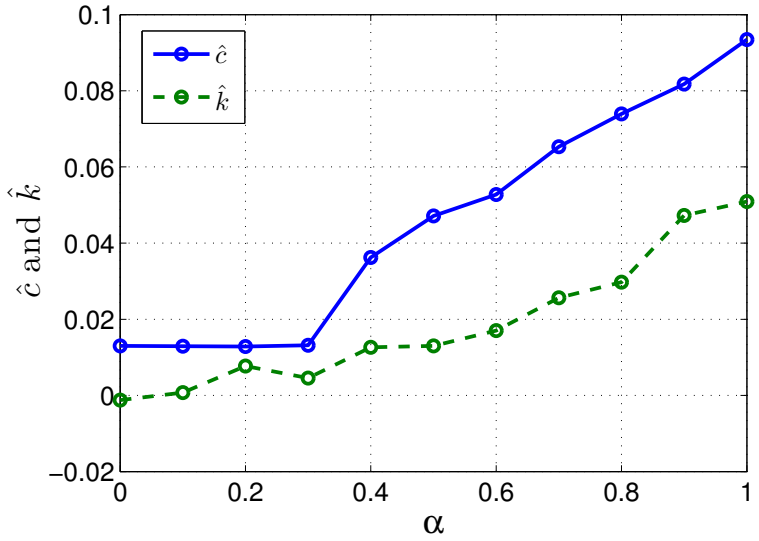


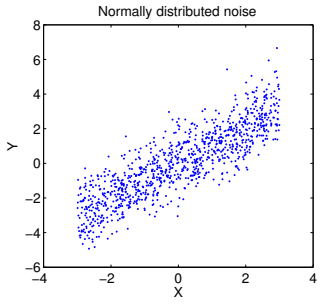
Fig. 1: The estimated parameters $\hat{c}$ and $\hat{k}$ in the functional causal model (17) when the true noise distribution in the data-generating process (15) varies from the Gaussian distribution (α=0) to the Exponential one (α=1).

4. ESTIMATING POST-NONLINEAR CAUSAL MODEL BY WARPED GAUSSIAN PROCESSES WITH A FLEXIBLE NOISE DISTRIBUTION

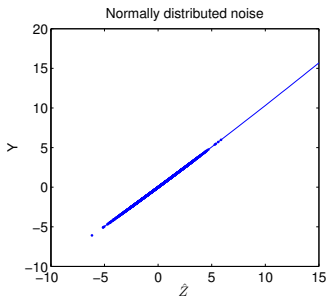
In this section we focus on the PNL causal model, since its form is very general and the causal direction is nevertheless identifiable in the general case (apart from the five special situations [Zhang and Hyvärinen 2009b]). It has been proposed to estimate the PNL causal model (2) by mutual information minimization [Zhang and Hyvärinen 2009b] with the nonlinear functions f_1 and f_2^{-1} represented by multi-layer preceptrons (MLPs). With this implementation, model selection for those nonlinear functions, i.e., selection of the numbers of hidden units in the MLPs, is non-trivial. With a too simple model, the estimated noise tends to be more dependent on the hypothetical cause, and a too complex one tends to cause over-fitting, such that the wrong causal direction could appear plausible. Moreover, the solution was found to be dependent on initializations of the nonlinear functions, i.e., it is prone to local optima.

As stated in Section 3, for any functional causal model, minimizing the mutual information between noise and the hypothetical cause is equivalent to maximum likelihood with a flexible model for the noise distribution; moreover, it was claimed that for estimation of the functional causal model where noise is not additive, especially the PNL causal model, the solution would be sensitive to the assumed noise distribution. Therefore, we propose an approach for estimating the PNL causal model based on Bayesian inference, which allows automatic model selection, and a flexible model for the noise distribution.

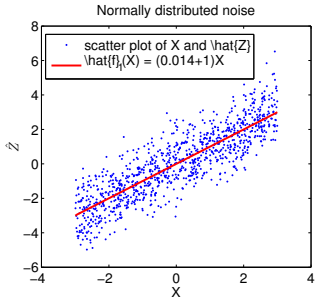
We adopt the warped Gaussian process [Snelson et al. 2004] framework, which can be interpreted as a two-step generative model of the output variable with values $\mathbf{y}_i \in \mathbb{R}$ given input variable with values $\mathbf{x}_i \in \mathbb{R}^d, i \in \{1, \ldots, n\}$, to specify the nonlinear functions and noise term in the PNL model (2). As stated in Section 3.2, for the PNL causal model, parameter estimation under a wrong noise model is not necessarily statistically consistent. Hence, a crucial difference between the original warped Gaussian pro-



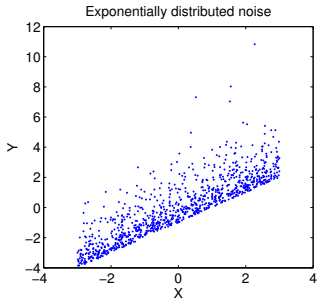
(a) X and Y data points (only 1000 points are shown) with *Gaussian* noise



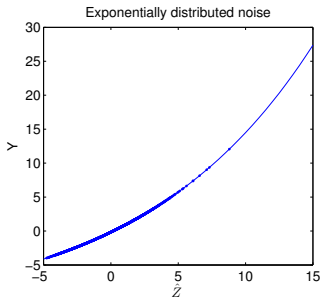
(b) Estimated f_2



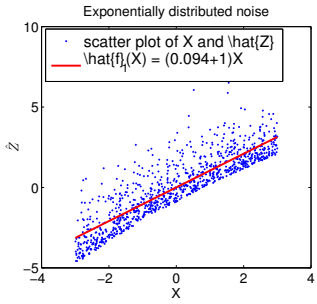
(c) X and $\hat{Z}$ as well as the estimated f_1



(d) X and Y data points (only 1000 points are shown) with *exponentially* distributed noise



(e) Estimated f_2



(f) X and $\hat{Z}$ as well as the estimated f_1

Fig. 2: Examples to illustrate that the estimated functional causal may be inconsistent with a wrongly specified noise model. The fitting functional causal model (17) assumes a standard Gaussian distribution for the noise term. Top: the true noise distribution is Gaussian. Bottom: the true noise distribution is an exponential one. From left to right: the (X, Y) data points (only 1000 are shown), and estimated function f_2, and the scatterplot of X and $\hat{Z}$ as well as the estimated f_1. Note that in the true data-generating process f_2 is an identity mapping.

cesses [Snelson et al. 2004] and our formulation is that the warped Gaussian process assumes Gaussian noise, but in our formulation the model for the noise distribution has to be flexible.

We will compare the performance of our proposed warped Gaussian process regression with the MoG noise (denoted by WGP-MoG) and that with the Gaussian noise (denoted by WGP-Gaussian) with simulations.

4.1. The model and prior

In the first step, an unknown function $f_1 : \mathbb{R}^d \to \mathbb{R}$ maps the value of the input variable, $\mathbf{x}_i$ to a latent variable

$$\mathbf{z}_i = f_1(\mathbf{x}_i) + \mathbf{n}_i, \tag{19}$$

where $\mathbf{n}_i \sim p(N; \Omega)$ is the noise distribution that is unknown. We approximate this noise distribution by a Mixture of Gaussian (MoG) distribution with parameters $\Omega =$

$\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$, given by

$$p(N|\Omega) = \sum_{j=1}^{m} \pi_j \mathcal{N}(N|\mu_j, \sigma_j^2), \tag{20}$$

where μ_j is the mean, σ_j the standard deviation, and π_j the positive mixing proportions that sum to one. We introduce latent membership variables $\phi_i \in \{1, \ldots, m\}$ that represent from which Gaussian components the noises ϵ_i were drawn. The membership variable ϕ_i follows a categorical distribution, i.e., $p(\phi_i = j|\Omega) = \pi_j$. In our implementation we set the number of Gaussian components $m = 5$.

We place a Gaussian process prior on the unknown function $f_1 \sim \mathcal{GP}(0, k(\cdot, \cdot))$ with a zero mean function. The GP is then fully determined by the covariance function $k(\cdot, \cdot)$. In this paper, we consider the isotropic Gaussian covariance function, given by

$$k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\Theta}) = \alpha_1 \exp\left(-\tfrac{\alpha_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \alpha_3 \delta_{\mathbf{x}_i, \mathbf{x}_j}, \tag{21}$$

with parameters $\boldsymbol{\Theta} = \{\alpha_1, \alpha_2, \alpha_3\}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two observations of the variable X.

Given the set of membership variables $\boldsymbol{\phi}$, the log posterior of the latent variables $\mathbf{z}$ is given by

$$\log p(\mathbf{z}|\mathbf{x}, \mathbf{C}, \Omega, \Theta) = -\frac{1}{2}\log\det(\mathbf{K} + \mathbf{C}) - \frac{1}{2}\bar{\mathbf{t}}^T(\mathbf{K} + \mathbf{C})^{-1}\bar{\mathbf{z}} - \frac{n}{2}\log(2\pi),$$

where $\mathbf{K}$ is the covariance matrix, i.e., $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{C}$ a diagonal noise variance matrix with $C_{i,i} = \sigma_{c_i}^2$, and $\bar{z}_i = z_i - \mu_{\phi_i}$ the latent variable subtracted by the noise mean.

In the second step, the latent variable $\mathbf{z}_i$ is mapped to the output space by function $f_2 : \mathbb{R} \to \mathbb{R}$, whose inverse is denoted by g, so we have

$$\mathbf{y}_i = g^{-1}(\mathbf{z}_i). \tag{22}$$

The post-nonlinear transformation in (2) represents the sensor distortion or measurement distortion; in practice, it is usually very smooth. We therefore use a rather simple representation for it. Following [Snelson et al. 2004], we choose the inverse warping function that is the sum of $\tanh$ functions and the identity function; for the ith value of Y, we have

$$g(\mathbf{y}_i; \boldsymbol{\Psi}) = \mathbf{y}_i + \sum_{i=1}^{k} a_i \tanh(b_i(\mathbf{y}_i + c_i)), \tag{23}$$

where the parameters $\boldsymbol{\Psi} = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and $a_i, b_i \geq 0, \forall i$, such that g is guaranteed to be strictly monotonic. Note that g^{-1} corresponds to f_2 in (2); for convenience of parameter estimation, here we directly parameterize f_2^{-1}, or g, instead of f_2.

Given the set of membership variables $\boldsymbol{\phi}$, the log posterior $\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}, \Omega, \Theta, \Psi)$ of the outputs $\mathbf{y}$ is given by

$$\mathcal{L}(\boldsymbol{\phi}) = -\frac{1}{2}\log\det(\mathbf{K} + \mathbf{C}) - \frac{1}{2}\bar{\mathbf{z}}^T(\mathbf{K} + \mathbf{C})^{-1}\bar{\mathbf{z}} + \sum_{i=1}^{n} \log \left.\frac{\partial g}{\partial y}\right|_{\mathbf{y}_i} - \frac{n}{2}\log(2\pi),$$

where $\bar{\mathbf{z}}_i = g(\mathbf{y}_i; \boldsymbol{\Psi}) - \mu_{\phi_i}$.

4.2. Parameter Learning

We use Monte Carlo Expectation Maximization [Levine and Casella 2001] to learn the parameters Ω, Θ, and Ψ, with the membership variables $\boldsymbol{\phi}$ marginalized out.

The Monte Carlo EM algorithm seeks to find the maximum likelihood estimate of the parameters by iteratively applying the following E-step and M-step.

In the E-step, we estimate

$$\mathcal{Q}(\Omega, \Theta, \Psi|\Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) = \mathbb{E}_{\boldsymbol{\phi}|\mathbf{x},\mathbf{y},\Omega^{(t)},\Theta^{(t)},\Psi^{(t)}}[\log \mathcal{L}(\boldsymbol{\phi})]. \tag{24}$$

However, direct evaluation of $\mathcal{Q}$ is intractable, since we do not have a parametric representation of posterior distribution of $\boldsymbol{\phi}$. We resort to approximating (24) with Gibbs sampling. That is, we calculate

$$\tilde{\mathcal{Q}}(\Omega, \Theta, \Psi|\Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) \triangleq \frac{1}{L}\sum_{l=1}^{L} \log \mathcal{L}(\boldsymbol{\phi}_l) \tag{25}$$

instead, where $\boldsymbol{\phi}_l$ is the lth value of $\boldsymbol{\phi}$ sampled from the posterior

$$p(\boldsymbol{\phi}|\mathbf{x}, \mathbf{y}, \Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}, \Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) p(\boldsymbol{\phi}|\Omega^{(t)}). \tag{26}$$

Here L is the total number of sampled values of $\boldsymbol{\phi}$.

In the M-step, we find the parameters $\Omega^{(t+1)}$, $\Theta^{(t+1)}$, and $\Psi^{(t+1)}$ that maximize the estimated $\tilde{\mathcal{Q}}(\Omega, \Theta, \Psi|\Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)})$ using scaled conjugate gradient.

The MATLAB source code for estimating the PNL causal model with WGP-MoG is available at http://people.tuebingen.mpg.de/kzhang/warpedGP.zip .

5. SIMULATION

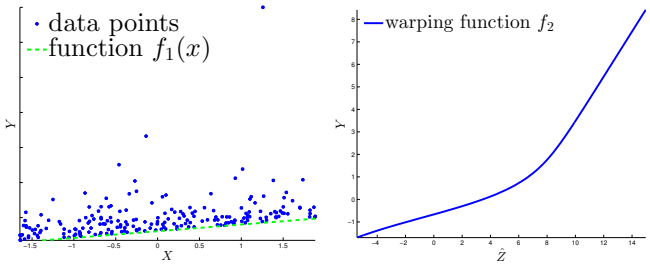
In this section we use simulated data to illustrate different behaviors of the proposed method for estimating the PNL causal model, which is based on warped Gaussian processes with noise represented by MoG, the original warped Gaussian process regression with the Gaussian noise [Snelson et al. 2004], and the mutual information minimization approach with nonlinear functions represented with MLPs [Zhang and Hyvärinen 2009b].[4] The linear additive noise model and the multiplicative noise model, both of which are special cases of the post-nonlinear causal model, were used for data generation.

5.1. Simulation 1: With Data Generated by a Linear Model

For illustrative purposes, we first use linear transformations for both f_1 and f_2 to see if they can be recovered by different methods. The one-dimensional inputs X were uniformly distributed; the latent variable $Z = f_1(X) + N$ were generated with a linear function $f_1(X) = 2X$, and the output $Y = f_2(Z)$ were generated with an identity warping function $f_2(Z) = Z$. The noise N were drawn from a log-normal distribution. We generated 200 data points. Figure 3a shows the simulated data points.

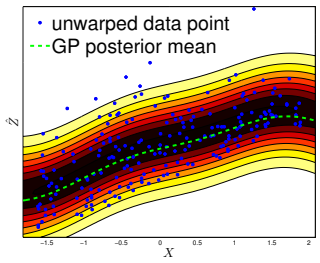
Figures 3 and 4 show the estimation results produced by WGP-Gaussian and WGP-MoG, respectively. One can see that in this case WGP-Gaussian gives clearly a wrong solution: the estimated post-nonlinear transformation f_2 is distorted in a specific way such that the estimated noise is closer to Gaussian that the true noise; as a consequence, the true data-generating process cannot be recovered by WGP-Gaussian, and finally the estimated noise is dependent from the input X, as seen from Figure 3d. With WGP-MoG, both estimated f_1 and f_2 were close to the true ones, which are actually linear. We increased the sample size to 500, and observed the same difference in

[4]As pointed out by an anonymous reviewer, for the implementation with MLPs, the difficulty in model selection of the MLPs could be addressed by exploiting a similar idea of using Gaussian process for hyperparameter optimization. In our implementation, we fixed the number of hidden units of the MLPs. Therefore. in this sense, the comparison between WGP-MoG and the MLP-based implementation conducted in this paper is not quite fair. We are grateful to the anonymous reviewer for making this clear.

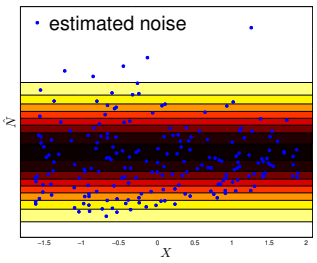


(a) Data

(b) Estimated PNL function f_2



(c) Estimated f_1



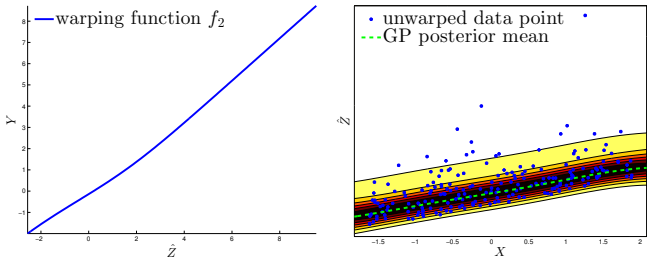
(d) X and estimated noise

Fig. 3: Simulation 1: Simulated data with the linear additive noise model (with log-normal noise) and estimation results by WGP-Gaussian. (a) Simulated data. (b) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (c) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$, where the dashed lines showed the GP posterior mean of $f_1(X)$, and the heat maps showed the conditional probability $p(\hat{Z}|X)$. (d) Scatter plot of input x_i and the estimated noise $\hat{N}_i$, where the heat maps showed the conditional probability $p(\hat{N}|X)$.

the estimated f_2 and f_1 given by WGP-MoG and WGP-Gaussian. This illustrates that the estimated f_2 and f_1 in the PNL causal model (2) might not be statistically consistent if the noise distribution is set to Gaussian incorrectly, and verifies the statement given in Section 3.2.

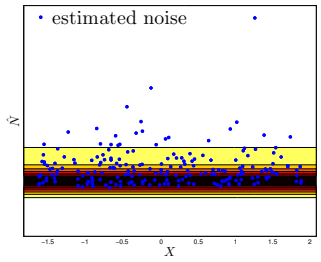
We also compare the above two approaches with mutual information minimization approach with nonlinear functions represented by MLPs [Zhang and Hyvärinen 2009b], whose results are shown in Figure 5. This approach also uses a MoG to represent the noise distribution, and could estimate both function f_1 and f_2, as well as the noise term, reasonably well in this simple situation.

We then distinguish cause from effect by estimating the PNL model followed by testing if the estimated noise is independent from the hypothetical cause for both directions. We adopted the Hilbert Schmidt information criterion (HSIC) [Gretton et al. 2008] for statistical independence test and set the significance level to $\alpha = 0.05$. Both WGP-MoG and the mutual information minimization approach correctly determined the causal direction, which is $X \to Y$, in that for $X \to Y$ the estimated noise is independent from X while for $Y \to X$ the estimated noise is dependent on Y. When using WGP-Gaussian, we found that the noise is dependent from the hypothetical cause for both directions with the significance level 0.05, although the p-value for the direction $X \to Y$ is larger (0.048 for $X \to Y$ and 0.010 for $Y \to X$).

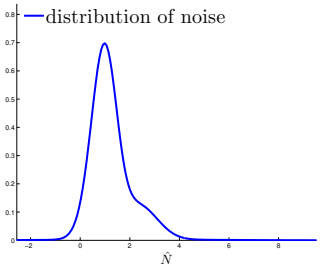


(a) Estimated PNL function f_2

(b) Estimated f_1



(c) X and estimated noise



(d) MoG Noise Distribution

Fig. 4: Simulation 1: Estimation results by WGP-MoG. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (b) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ using WGP-MoG, where the dashed lines showed the GP posterior mean of $f_1(X)$, and the heat maps showed the conditional probability $p(\hat{Z}|X)$. (c) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$, where the heat maps showed the conditional probability $p(\hat{N}|X)$. (d) Estimated noise distribution $p(\hat{N})$.

5.2. Simulation 2: With Data Generated by a Multiplicative Noise Model

In the second simulation, we generated 400 data point according to the multiplicative noise model,

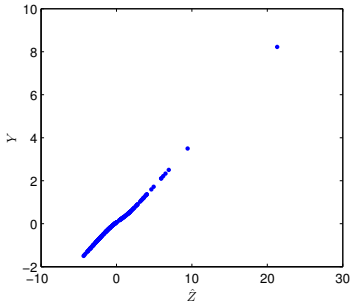
$$Y = X \cdot \mathring{N}, \tag{27}$$

where X takes positve values and is uniformly distributed, and $\mathring{N} = e^N$ with N being the absolute value of the standard Gaussian variable. Note that this model actually belongs to the class of post-nonlinear models: it can be rewritten as

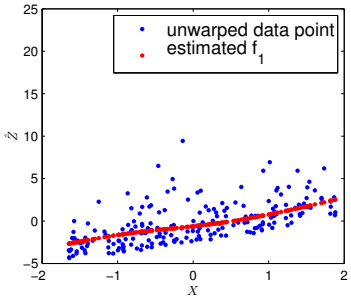
$$Y = e^{\log X + \log \mathring{N}} = e^Z = e^{\log X + N}, \tag{28}$$

and hence it is a special case of (2) with $f_1(X) = \log X$ and f_2 being the exponential transformation. Figures 6a shows the generated data points.

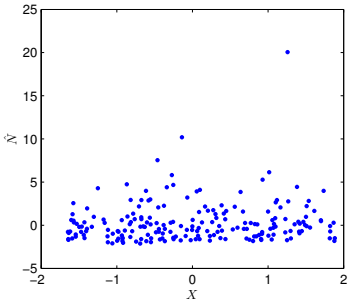
Figures 6 and 7 show the estimation results by WGP-Gaussian and WGP-MoG, respectively. According to (28), if the warping function f_2 is perfectly recovered, the recovered latent variable $\hat{Z}$ will have a linear relationship with $\log Y$. Comparing Fig. 6c with Fig. 7b, one can then see that WGP-MoG gives a better estimate of f_2. Furthermore, as shown by Fig. 6e and Fig. 7d (or Fig. 7e), the noise estimated by WGP-Gaussian becomes closer to Gaussian, and the distribution of the noise estimated by WGP-MoG is closer to the absolute value of the standard Gaussian (up to a location and scale transformation), which was the distribution of N in (28) for data generation. To save space, we skip the estimation results by MLPs-based mutual information



(a) Estimated PNL function f_2



(b) Estimated f_1



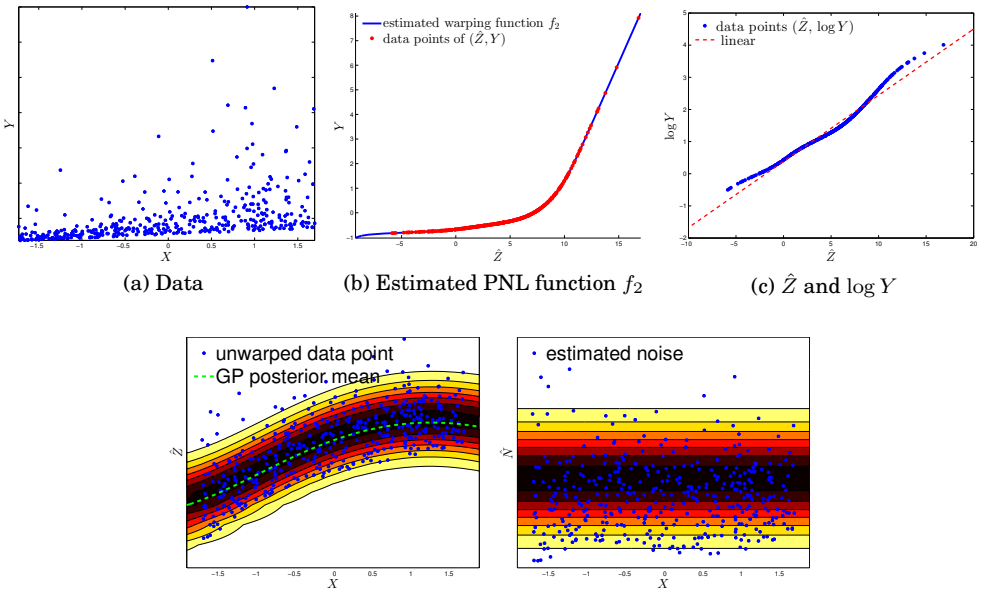
(c) X and estimated noise

Fig. 5: Simulation 1: Estimation results by mutual information minimization with nonlinear functions represented by MLPs. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (b) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ where the red points show $\hat{f}_1(\mathbf{x}_i)$. (c) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$

minimization approach and the results of distinguishing cause from effect on this data set.

6. ON REAL DATA

We applied different approaches for causal direction determination on the cause-effect pairs available at http://webdav.tuebingen.mpg.de/cause-effect/. The approaches include the PNL causal model estimated by mutual information minimization with nonlinear functions represented by MLPs [Zhang and Hyvärinen 2009b], denoted by PNL-MLP for short, the PNL causal model estimated by warped Gausian processes with Gaussian noise, denoted by PNL-WGP-Gaussian, the PNL causal model estimated by warped Gausian processes with MoG noise, denoted by PNL-WGP-MoG, the additive noise model estimated by Gaussian process regression [Hoyer et al. 2009], denoted by ANM, the approach based on the Gaussian process prior on the function f [Mooij et al. 2010], denoted by GPI, and IGCI [Janzing et al. 2012]. The data set consists of 77 data pairs. To reduce computational load, we used at most 500 points for each cause-effect pair: if the original data set consists of more than 500 points, we randomly sampled 500 points from them; otherwise we simply used the original data set. The accuracy of different methods (in terms of the percentage of correctly discovered causal directions)



(a) Data (b) Estimated PNL function f_2 (c) $\hat{Z}$ and $\log Y$

(d) Estimated f_1 (e) X and estimated noise

Fig. 6: Simulation 2: Simulated data with the multiplicative noise model (27) and estimation results by WGP-Gaussian. (a) Simulated data. (b) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (c) Scatterplot of the recovered latent variable $\hat{\mathbf{z}}_i$ and $\log \mathbf{y}_i$, which will be on a line if f_2 is perfectly recovered. (d) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$, where the dashed lines showed the GP posterior mean of $f_1(X)$, and the heat maps showed the conditional probability $p(\hat{Z}|X)$. (e) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{N}_i$, where the heat maps showed the conditional probability $p(\hat{N}|X)$.

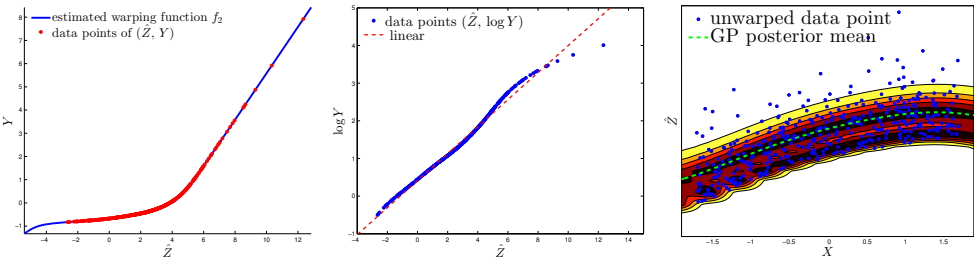
Table I: Accuracy of different methods for causal direction determination on the cause-effect pairs.

Method	PNL-MLP	PNL-WGP-Gaussian	PNL-WGP-MoG	ANM	GPI	IGCI
Accuracy (%)	70	67	76	63	72	73

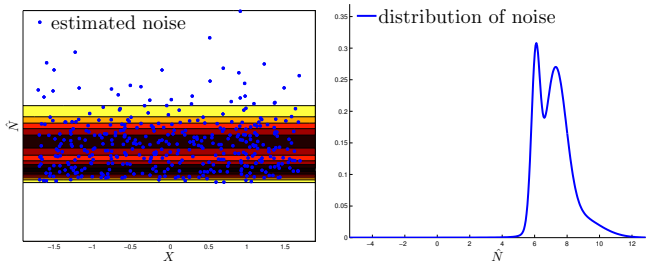
is reported in Table I. One can see that PNL-WGP-MoG gives the best performance among these methods.

On several data sets PNL-WGP-Gaussian and PNL-WGP-MoG give different conclusions. For instance, on both data pairs 22 and 57, PNL-WGP-Gaussian prefers $Y \to X$, and PNL-WGP-MoG prefers $X \to Y$, which would be the plausible one according to the background knowledge. In fact, for data pair 22, X corresponds to the age of a particular person, and Y is the corresponding height of the same person; for data pair 57, X denotes the latitude of the country's capital, and Y is the life expectancy at birth in the same country.

Figures 8 and 10 show the estimated post-nonlinear transformations f_2, functions f_1, and the noise N produced by PNL-WGP-Gaussian, under both hypothetical causal directions $X \to Y$ and $Y \to X$, on data pairs 22 and 57, respectively. For comparison,



(a) Estimated PNL function f_2 (b) $\hat{Z}$ and $\log Y$ (c) Estimated f_1



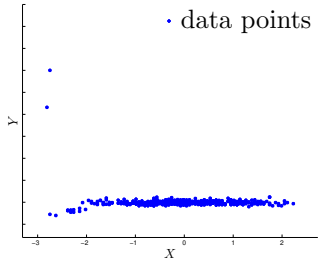
(d) X and estimated noise (e) MoG Noise Distribution

Fig. 7: Simulation 2: Estimation results by WGP-MoG. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (b) Scatterplot of the recovered latent variable $\hat{\mathbf{z}}_i$ and $\log \mathbf{y}_i$, which will be on a line if f_2 is perfectly recovered. (c) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ using WGP-MoG, where the dashed lines showed the GP posterior mean of $f_1(X)$, and the heat maps showed the conditional probability $p(\hat{Z}|X)$. (d) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$, where the heat maps showed the conditional probability $p(\hat{N}|X)$. (e) Estimated noise distribution $p(\hat{N})$.

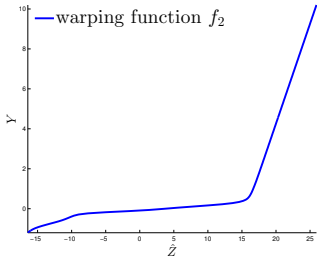
Figure 9 and 11 show the results produced by PNL-WGP-MoG on the two data sets. One can see that PNL-WGP-Gaussian tends to push the noise distribution closer to Gaussian, making the estimated noise tend to be more dependent on the hypothetical cause. Overall, PNL-WGP-MoG clearly outperforms PNL-WGP-Gaussian in terms of the estimation quality of the PNL causal model and the performance of causal direction determination.

7. CONCLUSION AND DISCUSSIONS

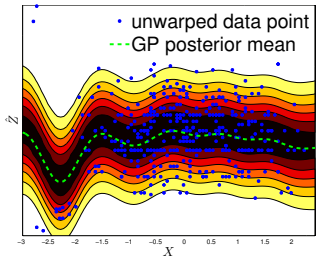
A functional causal model represents the effect as a function of the direct causes and a noise term which is independent from the direct causes. Suppose two given variables have a direct causal relation in between and that there is no confounder. A functional causal model could determine the causal direction between them if 1) it could approximate the true data-generating process, and 2) it holds for only one direction. When using functional causal models for causal direction determination, one has to find the direction in which the noise term is independent from the hypothetical cause. Under the hypothetical causal direction, a natural way to estimate the function and noise is to minimize the dependence between noise and the hypothetical cause. In this paper, we have shown that minimizing the mutual information between them is equivalent to maximizing the data likelihood if the model for the noise distribution is flexible. In this way, the two model estimation principles are unified.



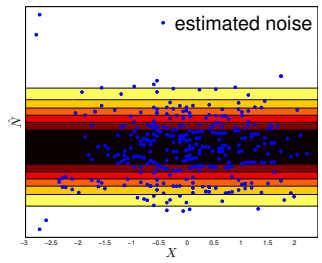
(a) Data pair 22



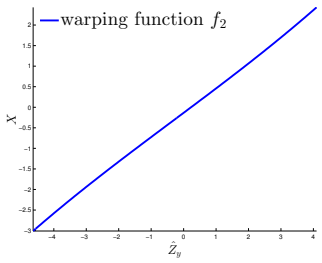
(b) Estimated PNL function f_2 for $X \to Y$



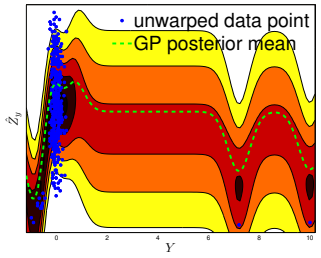
(c) Estimated f_1 for $X \to Y$



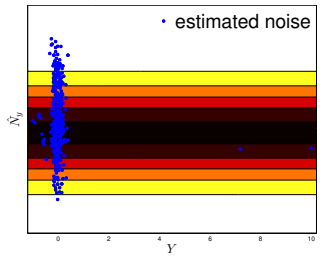
(d) X and estimated noise N for $X \to Y$



(e) Estimated PNL function f_2 for $Y \to X$



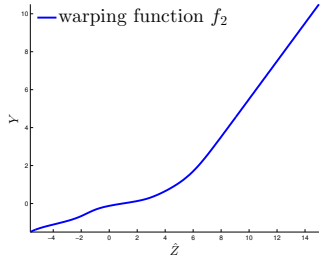
(f) Estimated f_1 for $Y \to X$



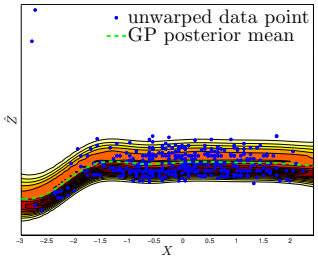
(g) Y and estimated noise N_Y for $Y \to X$

Fig. 8: Estimated PNL causal model for hypothetical causal direction $X \to Y$ (b-d) and direction $Y \to X$ (e-g) on cause-effect pair 22 by **PNL-WGP-Gaussian**. (a) Data. Here X and Y represent the age (in years) and height (in centimeters) of 452 patients, so it is plausible to have $X \to Y$. (b) Estimated warping function $Y = \hat{f}_2(\hat{Z})$ under $X \to Y$. (c) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ under $X \to Y$. (d) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$ under $X \to Y$, with the p-value of the HSIC independence test 0.0070. (e) Estimated warping function $\hat{f}_2$ under $Y \to X$. (f) Scatter plot of input $\mathbf{y}_i$ and the recovered latent variable $\hat{f}_2^{-1}(\mathbf{y}_i)$ under $Y \to X$. (g) Scatter plot of input $\mathbf{y}_i$ and the estimated noise $\hat{\mathbf{n}}_{Y,i}$ under $Y \to X$, with the p-value of the HSIC independence test 0.0470.

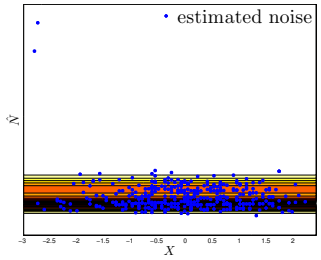
Furthermore, we have discussed that for a general functional causal model where noise is not additive, estimation of the function as well as the noise term might not be statistically consistent if the noise model is wrong. In light of these two points, we advocate the Bayesian inference based approach with a flexible noise model to



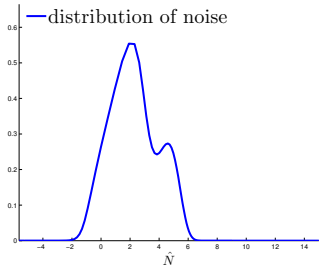
(a) Estimated PNL function f_2 for $X \to Y$



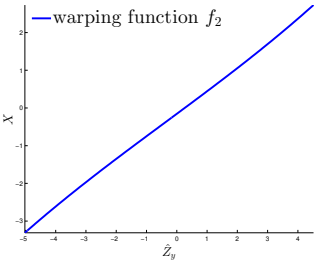
(b) Estimated f_1 for $X \to Y$



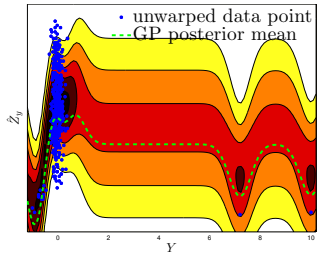
(c) X and estimated noise N for $X \to Y$



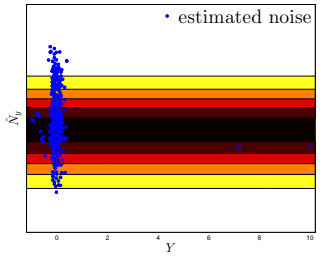
(d) MoG Noise Distribution under $X \to Y$



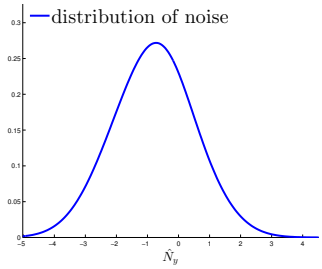
(e) Estimated PNL function f_2 for $Y \to X$



(f) Estimated f_1 for $Y \to X$



(g) Y and estimated noise N_Y for $Y \to X$

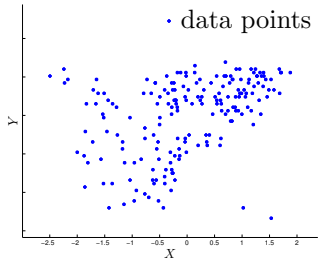


(h) MoG Noise Distribution under $Y \to X$

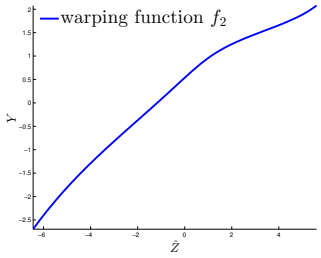
Fig. 9: Estimated PNL causal model for hypothetical causal direction $X \to Y$ (a-d) and direction $Y \to X$ (e-h) on cause-effect pair 22 by **PNL-WGP-MoG**. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$ under $X \to Y$. (b) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ under $X \to Y$. (c) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$ under $X \to Y$, with the p-value of the HSIC independence test 0.3090. (d) Estimated noise distribution $p(\hat{N})$ under $X \to Y$. (e) Estimated warping function $\hat{f}_2$ under $Y \to X$. (f) Scatter plot of input $\mathbf{y}_i$ and the recovered latent variable $\hat{f}_2^{-1}(\mathbf{y}_i)$ under $Y \to X$. (g) Scatter plot of input $\mathbf{y}_i$ and the estimated noise $\hat{\mathbf{n}}_{Y,i}$ under $Y \to X$, with the p-value of the HSIC independence test 0.0480. (h) Estimated noise distribution $p(\hat{N}_Y)$ under $Y \to X$.

estimation of functional causal models of a more general form than the additive noise model.

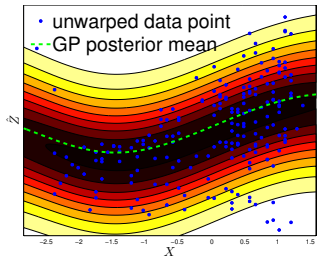
In particular, we focused on estimation of the post-nonlinear causal model, and proposed to estimate it by warped Gaussian processes with the noise distribution repre-



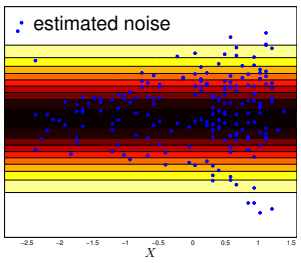
(a) Data pair 57



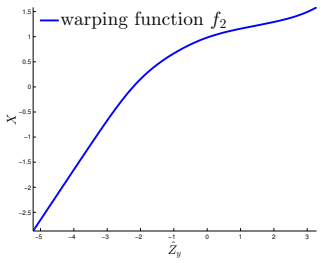
(b) Estimated PNL function f_2 for $X \to Y$



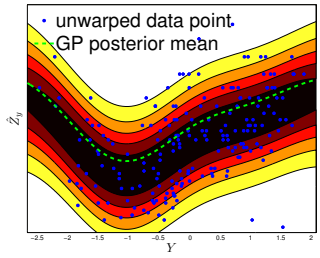
(c) Estimated f_1 for $X \to Y$



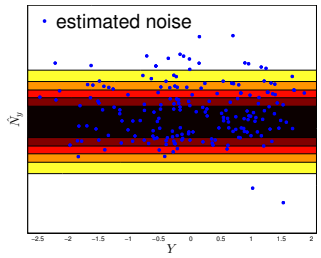
(d) X and estimated noise N for $X \to Y$



(e) Estimated PNL function f_2 for $Y \to X$



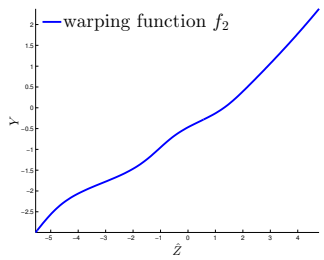
(f) Estimated f_1 for $Y \to X$



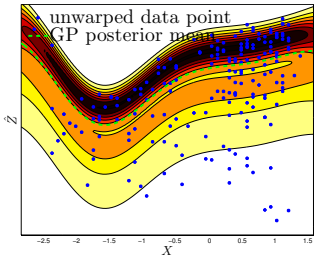
(g) Y and estimated noise N_Y for $Y \to X$

Fig. 10: Estimated PNL causal model for hypothetical causal direction $X \to Y$ (b-d) and direction $Y \to X$ (e-g) on cause-effect pair 57 by **PNL-WGP-Gaussian**. (a) Data. Here X and Y represent the latitude of each country's capital and the female life expectancy at birth for the same country. Naturally one would prefer $X \to Y$. (b) Estimated warping function $Y = \hat{f}_2(\hat{Z})$ under $X \to Y$. (c) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ under $X \to Y$. (d) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$ under $X \to Y$, with the p-value of the HSIC independence test 0.0220. (e) Estimated warping function $\hat{f}_2$ under $Y \to X$. (f) Scatter plot of input $\mathbf{y}_i$ and the recovered latent variable $\hat{f}_2^{-1}(\mathbf{y}_i)$ under $Y \to X$. (g) Scatter plot of input $\mathbf{y}_i$ and the estimated noise $\hat{\mathbf{n}}_{Y,i}$ under $Y \to X$, with the p-value of the HSIC independence test 0.0920.

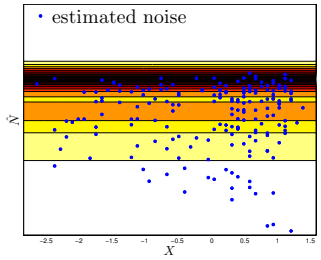
sented by the mixture of Gaussians. We exploited Monte Carlo EM for inference and parameter learning. Experimental results on simulated data illustrated that when the noise distribution is far from Gaussian, this approach is able to recover the data-



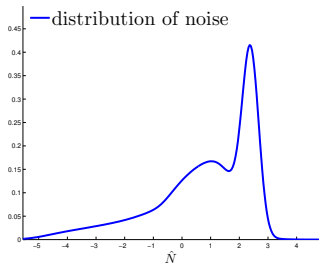
(a) Estimated PNL function f_2 for $X \to Y$



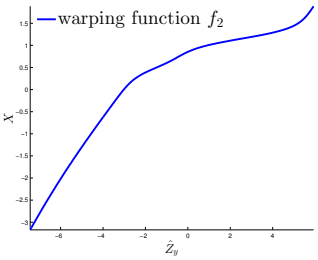
(b) Estimated f_1 for $X \to Y$



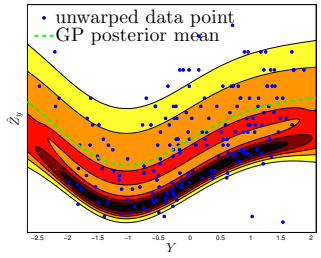
(c) X and estimated noise N for $X \to Y$



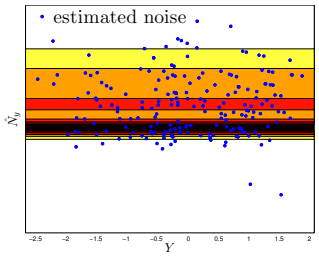
(d) MoG Noise Distribution under $X \to Y$



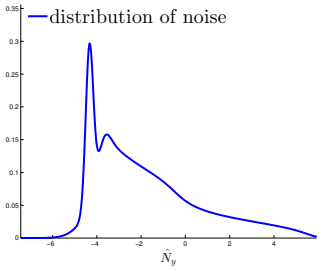
(e) Estimated PNL function f_2 for $Y \to X$



(f) Estimated f_1 for $Y \to X$



(g) Y and estimated noise N_Y for $Y \to X$



(h) MoG Noise Distribution under $Y \to X$

Fig. 11: Estimated PNL causal model for hypothetical causal direction $X \to Y$ (a-d) and direction $Y \to X$ (e-h) on cause-effect pair 57 by **PNL-WGP-MoG**. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$ under $X \to Y$. (b) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ under $X \to Y$. (c) Scatter plot of input $\mathbf{x}_i$ and the estimated noise $\hat{\mathbf{n}}_i$ under $X \to Y$, with the p-value of the HSIC independence test 0.8540. (d) Estimated noise distribution $p(\hat{N})$ under $X \to Y$. (e) Estimated warping function $\hat{f}_2$ under $Y \to X$. (f) Scatter plot of input $\mathbf{y}_i$ and the recovered latent variable $\hat{f}_2^{-1}(\mathbf{y}_i)$ under $Y \to X$. (g) Scatter plot of input $\mathbf{y}_i$ and the estimated noise $\hat{\mathbf{n}}_{Y,i}$ under $Y \to X$, with the p-value of the HSIC independence test 0.1420. (h) Estimated noise distribution $p(\hat{N}_Y)$ under $Y \to X$.

generating process as well as the noise distribution, while the warped Gaussian processes with the Gaussian noise could fail. We used the proposed approach to estimation of the post-nonlinear causal model for determining causal directions on real data, and the experimental results showed that the proposed approach outperforms other meth-

ods for estimating the post-nonlinear causal model and other state-of-the-art methods for causal direction determination.

Finally, we would like to remark that the background causal knowledge has been demonstrated to be able to facilitate understanding and solving some machine learning problems, including semi-supervised learning [Schölkopf et al. 2012] and domain adaptation [Zhang et al. 2013a]. In these scenarios one does not aim to find causal directions, but may need to estimate the transformation from the cause to the effect, and the theoretical results given in this paper, such as Lemma 1 and the discussion in Section 3.2, might also help.

Acknowledgement

The research of J. Zhang was supported in part by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

REFERENCES

P. J. Bickel and K. A. Doksum. 1981. An analysis of transformations revisited. *J. Amer. Statist. Assoc.* 76 (1981), 296–311.

T. M. Cover and J. A. Thomas. 1991. *Elements of Information Theory*. Wiley.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. 2008. A Kernel Statistical Test of Independence. In *NIPS 20*. MIT Press, Cambridge, MA, 585–592.

P.O. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. 2009. Nonlinear Causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*. Vancouver, B.C., Canada.

A. Hyvärinen, J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. John Wiley & Sons, Inc.

A. Hyvärinen and P. Pajunen. 1999. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks* 12, 3 (1999), 429–439.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniuvsis, B. Steudel, and B. Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* (2012), 1–31.

R. A. Levine and G. Casella. 2001. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* 10, 3 (2001), 422–439.

J. Mooij, Janzing D., J. Peters, and B. Schölkopf. 2009. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML2009)*. 745–752.

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

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

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. 2012. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning (ICML 2012)*. Edinburgh, Scotland.

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

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. 2004. Warped Gaussian Processes. In *Advances in Neural Information Processing Systems 16*. MIT Press.

P. Spirtes, C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search* (2nd ed.). MIT Press, Cambridge, MA.

A. Taleb and C. Jutten. 1999. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing* 47, 10 (1999), 2807–2820.

M. Yamada and M. Sugiyama. 2010. Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI-2010)*. 643–648.

K. Zhang and L. Chan. 2005. Extended Gaussianization method for blind separation of post-nonlinear mixtures. *Neural Computation* 17, 2 (2005), 425–452.

K. Zhang and A. Hyvärinen. 2009a. Acyclic causality discovery with additive noise: An information-theoretical perspective. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2009*. Bled, Slovenia.

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

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2011. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. Barcelona, Spain.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. 2013a. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, JMLR: W&CP Vol. 28*.

K. Zhang, Z. Wang, and B. Schölkopf. 2013b. On Estimation of Functional Causal Models: Post-Nonlinear Causal Model as an Example. In *Proceedings of IEEE 13th International Conference on Data Mining Workshops*. Dallas, US, 139–146.