On Estimation of Functional Causal Models: Post-Nonlinear Causal Model as an Example

Kun Zhang, Zhikun Wang, Bernhard Schölkopf
Dept. Empirical Inference
Max-Planck Institute for Intelligent Systems
72076 Tübingen, Germany
Email: {kzhang, zhikun, bs}@tuebingen.mpg.de

***Abstract*—Compared to constraint-based causal discovery, causal discovery based on functional causal models is able to identify the whole causal model under appropriate assumptions. 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.**

I. 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 observed 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 [1], [2]. This approach involves the conditional independence test, which would be a difficult task if the data dependence relationship is unknown [3]. 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:

$$Y = f(X, N; \boldsymbol{\theta}_1), \tag{1.1}$$

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 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 [4]), f is linear, and at most one of the noise N and cause X is Gaussian. The nonlinear additive noise model [5], [6] assumes that f is nonlinear with additive noise N. In the post-nonlinear (PNL) causal model [7], the effect Y is further generated by a post-nonlinear transformation on the nonlinear effect of the cause plus the noise; the post-nonlinear transformation could represent the sensor distortion 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 [7]).

For causal discovery based on the nonlinear additive noise model, some regression methods have been proposed to directly minimize the dependence between the noise and the hypothetical cause [8], [9]. 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 the noise is not necessarily additive, minimizing the mutual information between the noise and the predictor is equivalent to maximizing the data likelihood, given that the noise model is flexible. As the second contribution, we show that for estimation of the functional causal model where the 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. In particular, as the third contribution, we finally focus on the PNL causal model, and 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) [7].

II. Asymmetry of Cause and Effect in Functional Causal Models

In this section we explain why f in the functional causal model (1.1) has to be properly constrained, and then give some examples of the functional forms f, including the PNL causal model.

A. 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, the quantity $\tilde{N} = q \circ F_{Y|X}$, where $F_{Y|X}$ is the conditional cumulative distribution function of Y given X and q is an arbitrary continuous and strictly monotonic function with a non-zero derivative, is always independent from X. Furthermore, the transformation from $(X, Y)^T$ to $(X, \tilde{N})^T$ is always invertible.

Proof: See [10] for why $\tilde{N}$ constructed this way is independent from X. 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 $q' \cdot p(Y|X)$, 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 (1.1), 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 the noise and hypothetical cause holds for only one direction.

B. 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 the 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 ([4]); 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 in taken into account in the functional class. A very general setting for $\mathcal{F}$ is given by the PNL causal model [7]:

$$Y = f_2(f_1(X) + N), \tag{2.2}$$

where both f_1 and f_2 are nonlinear functions and f_2 is assumed to be invertible. The post-nonlinear transformation f_2 could represent sensor distortion or measurement distortion in the system. It has been shown that 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 [5].

III. 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; for simplicity here we assume that X is one-dimensional and that there is no common cause for X and Y. The same result will also apply if X contains multiple variables.

We consider the functional causal model (1.1). Denote by $p(X, Y)$ the true density of (X, Y), and by $p_{\mathcal{F}}(X, Y)$ the joint density implied by (1.1). The model (1.1) assumes $p(X, N) = p(X)p(N)$; because the Jacobian matrix of the transformation from $(X, N)^T$ to $(X, Y)^T$ is

$$\mathbf{J}_{X\to Y} = \begin{pmatrix} \frac{\partial X}{\partial X} & \frac{\partial X}{\partial N} \\ \frac{\partial Y}{\partial X} & \frac{\partial Y}{\partial N} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{\partial f}{\partial X} & \frac{\partial f}{\partial N} \end{pmatrix}, \tag{3.3}$$

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

$$P_{\mathcal{F}}(X, Y) = p(X, N)/|\mathbf{J}_{X\to Y}| = p(X)p(N)\left|\frac{\partial f}{\partial N}\right|^{-1}, \tag{3.4}$$

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

Now let us introduce the concept of mutual information ([11]). As a canonical measure of statistical dependence, mutual information between X and N is defined as:

$$\begin{aligned} I(X, N) &= \int p(X, N) \log \frac{p(X, N)}{p(X)p(N)} dxdn \\ &= -\mathbb{E} \log p(X) - \mathbb{E} \log p(N) + \mathbb{E} \log p(X, N), \end{aligned} \tag{3.5}$$

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

A. Maximum likelihood and dependence minimization for functional causal models

Suppose we fit the model (1.1) 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 first show that the attained likelihood of (1.1) is directly related to the dependence between the estimated noise N and X. We further denote by $\boldsymbol{\theta}_2$ the parameter set in $p(N)$.

LEMMA 3.1: For any parameter set $\boldsymbol{\theta} \triangleq (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, the log-likelihood attained by the model (1.1) is

$$\begin{aligned} l_{X\to Y}(\boldsymbol{\theta}) &= \sum_{i=1}^{T} \log P_{\mathcal{F}}(\mathbf{x}_i, \mathbf{y}_i) \\ &= \sum_{i=1}^{T} \log p(X = \mathbf{x}_i) + \sum_{i=1}^{T} \log p(N = \hat{\mathbf{n}}_i; \boldsymbol{\theta}_2) \\ &\quad - \sum_{i=1}^{T} \log \Big| \frac{\partial f}{\partial N} \big|_{N=\hat{\mathbf{n}}_i} \Big|. \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} I(X, \hat{N}; \boldsymbol{\theta}) &= -\frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i) - \frac{1}{T}\sum_{i=1}^{T} \log p(\hat{N} = \hat{\mathbf{n}}_i; \boldsymbol{\theta}_2) \\ &+ \frac{1}{T}\sum_{i=1}^{T} \log \Big| \frac{\partial f}{\partial N} \big|_{N=\hat{\mathbf{n}}_i} \Big| + \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i), \end{aligned} \tag{3.7}$$

where the last term does not depend on $\boldsymbol{\theta}$ and can then be considered as constant.

Hence, for any value of the parameter set $\boldsymbol{\theta}$, we have

$$\frac{1}{T} l_{X\to Y}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(X, \hat{N}; \boldsymbol{\theta}). \tag{3.8}$$

Therefore, the parameter set $\boldsymbol{\theta}^*$ that maximizes the likelihood of the model (1.1) also minimizes the mutual information $I(X, \hat{N})$.

Proof: (3.6) directly follows (3.4), and now we prove (3.7). 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 (3.5), one can see

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

whose sample version is (3.7). (3.8) can be directly seen from (3.6) and (3.7). ■

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

$$X = g(Y, N_Y; \boldsymbol{\psi}) \tag{3.9}$$

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., (1.1) holds), but the reverse model (3.9) does not hold, i.e., these does not exist $g \in \mathcal{F}$ such that N_Y is independent from Y in (3.9), 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 1: Assume that the model (1.1) is true and that $\boldsymbol{\theta}^*$ maximizes the likelihood of the model (1.1) on the given sample $\mathcal{D}$. Further assume that $\boldsymbol{\psi}^*$ maximizes the likelihood of the model (3.9), but at $T \to \infty$, the resulting noise $\hat{N}_Y$ is dependent on Y, i.e., the model (3.9) does not hold. 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

$$\frac{1}{T} l_{X\to Y}(\boldsymbol{\theta}^*) - \frac{1}{T} l_{Y\to X}(\boldsymbol{\psi}^*) = I(Y, \hat{N}_Y; \boldsymbol{\psi}^*). \tag{3.10}$$

Proof: According to (3.8), we have

$$\frac{1}{T} l_{X\to Y}(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(X, \hat{N}; \boldsymbol{\theta}^*), \tag{3.11}$$

$$\frac{1}{T} l_{Y\to X}(\boldsymbol{\psi}^*) = \frac{1}{T}\sum_{i=1}^{T} \log p(X = \mathbf{x}_i, Y = \mathbf{y}_i) - I(Y, \hat{N}_Y; \boldsymbol{\psi}^*). \tag{3.12}$$

Bearing in mind that $I(X, \hat{N}; \boldsymbol{\theta}^*) \to 0$ as $T \to \infty$, one substracts (3.12) from (3.11) and obtains (3.10). ■

B. Loss Caused by a Wrongly Epecified Noise Distribution

As claimed in Lemma 3.1, estimating the functional causal model by maximum likelihood or mutual information minimization aims to maximize

$$J_{X\to Y} = \sum_{i=1}^{T} \log p(N = \hat{\mathbf{n}}_i) - \sum_{i=1}^{T} \log \Big| \frac{\partial f}{\partial N} \big|_{N=\hat{\mathbf{n}}_i} \Big|. \tag{3.13}$$

In the linear model (i.e., f in (1.1) 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.

This might not be the case for the general functional causal models. In fact, Bickel and Doksum [12] investigated the statistical consistency properties of the parameters in the Box-Cox transformation, which is a special case of the PNL formulation (2.2) 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 (2.2), 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 sacrifice in order to make the estimated noise closer to the specified distribution such that the first term in (3.13) becomes bigger; consequently, (3.13), a trade-off of the two terms, is maximized.

IV. 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 [7]). It has been proposed to estimate the PNL causal model (2.2) by mutual information minimization [7] with the nonlinear functions f_1 and f_2^{-1} represented by multilayer preceptrons (MLPs). This implementation suffers two drawbacks. First, it is difficult to do model selection for those nonlinear functions, i.e., selection of the numbers of hidden units in the MLPs; here 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 considered causal direction could be incorrectly plausible. Second, 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 III, for any functional causal model, minimizing the mutual information between the noise and the hypothetical cause is equivalent to minimum likelihood with a flexible model for the noise distribution; moreover, it was claimed that for estimation of the functional causal model where the 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 [13] 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 (2.2). As stated in Section III-B, 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 processes [13] 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.

A. 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

$$\mathbf{z}_i = f_1(\mathbf{x}_i) + \mathbf{n}_i, \tag{4.14}$$

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

$$p(N|\Omega) = \sum_{j=1}^{m} \pi_j \mathcal{N}(N|\mu_j, \sigma_j^2), \tag{4.15}$$

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 $\theta_i \in \{1, \ldots, m\}$ that represent from which Gaussian components the noises ϵ_i were drawn. The membership variable θ_i follows a categorical distribution, i.e., $p(\theta_i = j|\Omega) = \pi_j$. 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

$$k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\Theta}) = \alpha_1 \exp\left(-\tfrac{\alpha_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \alpha_3 \delta_{\mathbf{x}_i, \mathbf{x}_j}, \tag{4.16}$$

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 $\boldsymbol{\theta}$, 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 $\bar{z}_i = z_i - \mu_{\theta_i}$ 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

$$\mathbf{y}_i = g^{-1}(\mathbf{z}_i). \tag{4.17}$$

The post-nonlinear transformation in (2.2) represents the sensor distortion or measurement distortion; in practice, it is usually very smooth. We therefore use a rather simple representation for it. Following [13], 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

$$g(\mathbf{y}_i; \boldsymbol{\Psi}) = \mathbf{y}_i + \sum_{i=1}^{k} a_i \tanh(b_i(\mathbf{y}_i + c_i)), \tag{4.18}$$

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 (2.2); for convenience of parameter estimation, here we directly parameterize f_2^{-1}, or g, instead of f_2.

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

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

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

B. Parameter Learning

We use Monte Carlo Expectation Maximization [14] to learn the parameters Ω, Θ, and Ψ, with the membership variables $\boldsymbol{\theta}$ 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

$$\mathcal{Q}(\Omega, \Theta, \Psi|\Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x},\mathbf{y},\Omega^{(t)},\Theta^{(t)},\Psi^{(t)}}[\log \mathcal{L}(\boldsymbol{\theta})]. \tag{4.19}$$

However, the computation of $\mathcal{Q}$ is intractable. We resort to estimating

$$\tilde{\mathcal{Q}}(\Omega, \Theta, \Psi|\Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) = \frac{1}{L} \sum_{l=1}^{L} \log \mathcal{L}(\boldsymbol{\theta}_l) \tag{4.20}$$

by sampling $\boldsymbol{\theta}_l$ from

$$p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \Omega^{(t)}, \Theta^{(t)}, \Psi^{(t)})p(\boldsymbol{\theta}), \tag{4.21}$$

using Gibbs sampling.

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.

V. Simulation

We use simple simulations to illustrate the different behaviors of the proposed method for estimating the PNL causal model, which is based on warped Gaussian processes with the noise represented by MoG, the original warped Gaussian process regression with the Gaussian noise [13], and the mutual information minimization approach with nonlinear functions represented with MLPs [7].

The simulated data set consisted of 200 data points. The one-dimensional inputs X were uniformly distributed. For illustrative purposes, we use linear transformations for both f_1 and f_2 to see if they can be recovered by different methods: 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. Figure 1a shows the simulated data points.

Figures 1 and 2 show the estimated 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 1d. 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 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 (2.2) might not be statistically consistent if the noise distribution is set to Gaussian incorrectly.

We also compare the above two approaches with mutual information minimization approach with nonlinear functions represented by MLPs [7], whose results are shown in Figure 3. 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 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) [15] 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 $Y \to X$).

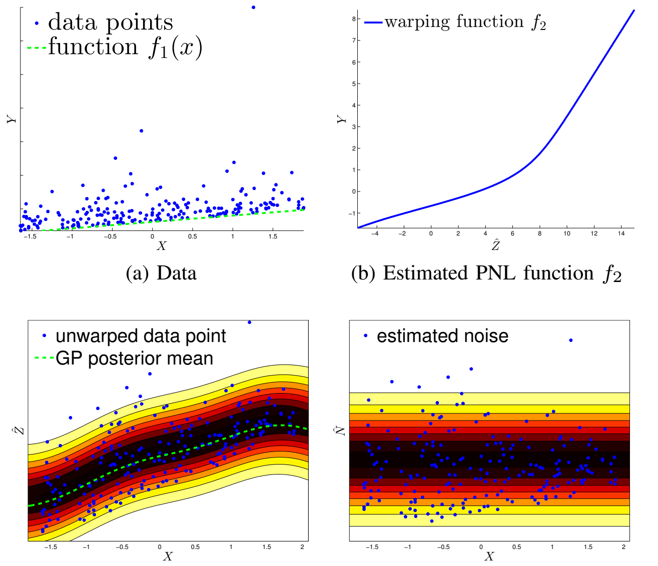
VI. On Real Data

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 [7], 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 [5], denoted by ANM, the approach based on the Gaussian process prior on the function f [16], denoted by GPI, and IGCI [17]. The data set consists of 77 data pairs. To reduce computational load, we used at most 500 points for each cause-effect pair. The accuracy of different methods (in terms of the percentage of correctly discovered causal directions) 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.

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



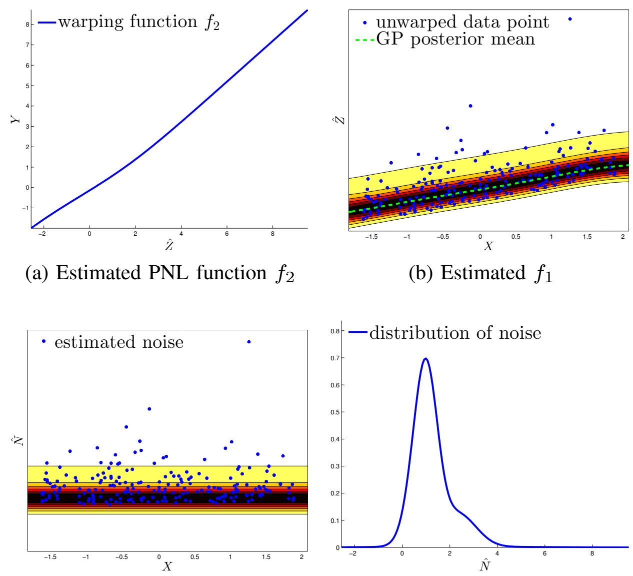
(a) Data

(b) Estimated PNL function f_2

(c) Estimated f_1

(d) X and estimated noise

Fig. 1: Simulated data with log-normal distributed 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 $p(\hat{N}|X)$.



(a) Estimated PNL function f_2

(b) Estimated f_1

(c) X and estimated noise

(d) MoG Noise Distribution

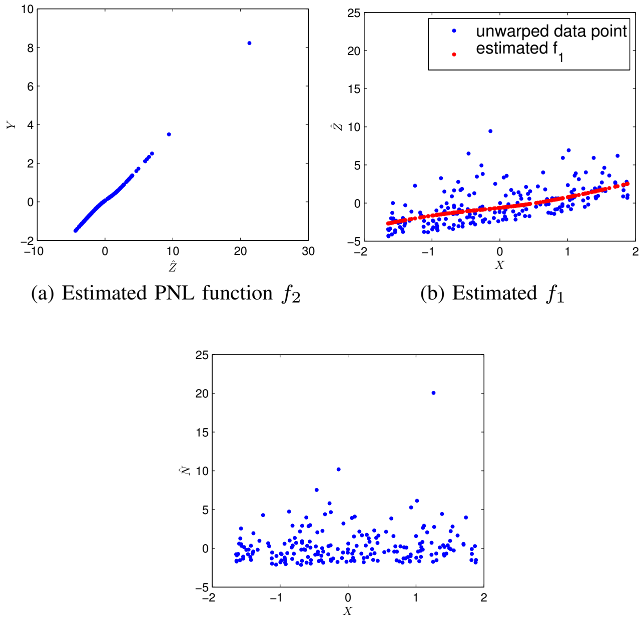
Fig. 2: Simulation results by WGP-MoG. (a) Estimated warping function $Y = \hat{f}_2(\hat{Z})$. (b) Scatter plot of input $\mathbf{x}_i$ and the recovered latent variable $\hat{\mathbf{z}}_i = \hat{f}_2^{-1}(\mathbf{y}_i)$ using WGP-MoG, where the dashed lines showed the GP posterior mean of $f_1(X)$, 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 $p(\hat{N})$.

Let us take data pair 22 as an example. Figures 4 shows the estimated post-nonlinear transformations f_2, functions f_1, and the noise N produced by PNL-WGP-Gaussian, under both hypothetical causal directions $X \rightarrow Y$ and $Y \rightarrow X$, on this data set. For comparison, Figure 5 gives the results produced by PNL-WGP-MoG on the same data set. 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.

VII. 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 the noise and 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. Furthermore, we have discussed that for a general functional causal model where the noise is not additive, estimation of the function as well as the noise 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 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 represented 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

(b) Estimated f_1

(c) X and estimated noise

Fig. 3: Simulation 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 $\hat{\mathbf{n}}_i$

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

Acknowledgment: ZK would like to acknowledge partial support of this research by the Research Grants Council of Hong Kong under the General Research Fund LU342213.

References

[1] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press, 2001.

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

[3] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "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, 2011.

[4] S. Shimizu, P. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, pp. 2003–2030, 2006.

[5] P. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Advances in Neural Information Processing Systems 21*, Vancouver, B.C., Canada, 2009.

[6] K. Zhang and A. Hyvärinen, "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, 2009.

[7] ——, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

[8] J. Mooij, J. D., J. Peters, and B. Schölkopf, "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)*, 2009, pp. 745–752.

[9] M. Yamada and M. Sugiyama, "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)*, 2010, pp. 643–648.

[10] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, 1999.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[12] P. J. Bickel and K. A. Doksum, "An analysis of transformations revisited," *Journal of the American Statistical Association*, vol. 76, pp. 296–311, 1981.

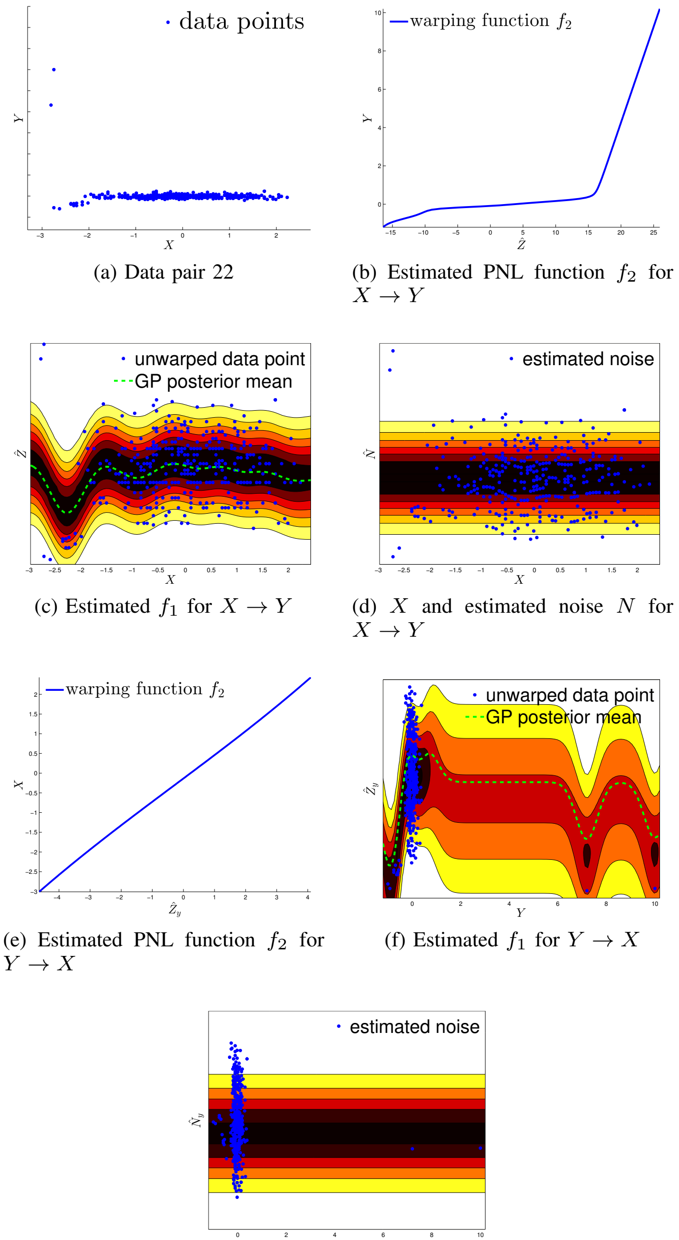
[13] E. Snelson, C. E. Rasmussen, and Z. Ghahramani, "Warped Gaussian processes," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[14] R. A. Levine and G. Casella, "Implementations of the Monte Carlo EM algorithm," *Journal of Computational and Graphical Statistics*, vol. 10, no. 3, pp. 422–439, 2001.

[15] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, "A kernel statistical test of independence," in *NIPS 20*. Cambridge, MA: MIT Press, 2008, pp. 585–592.

[16] J. 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)*, Curran, NY, USA, 2010.

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



(a) Data pair 22

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

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

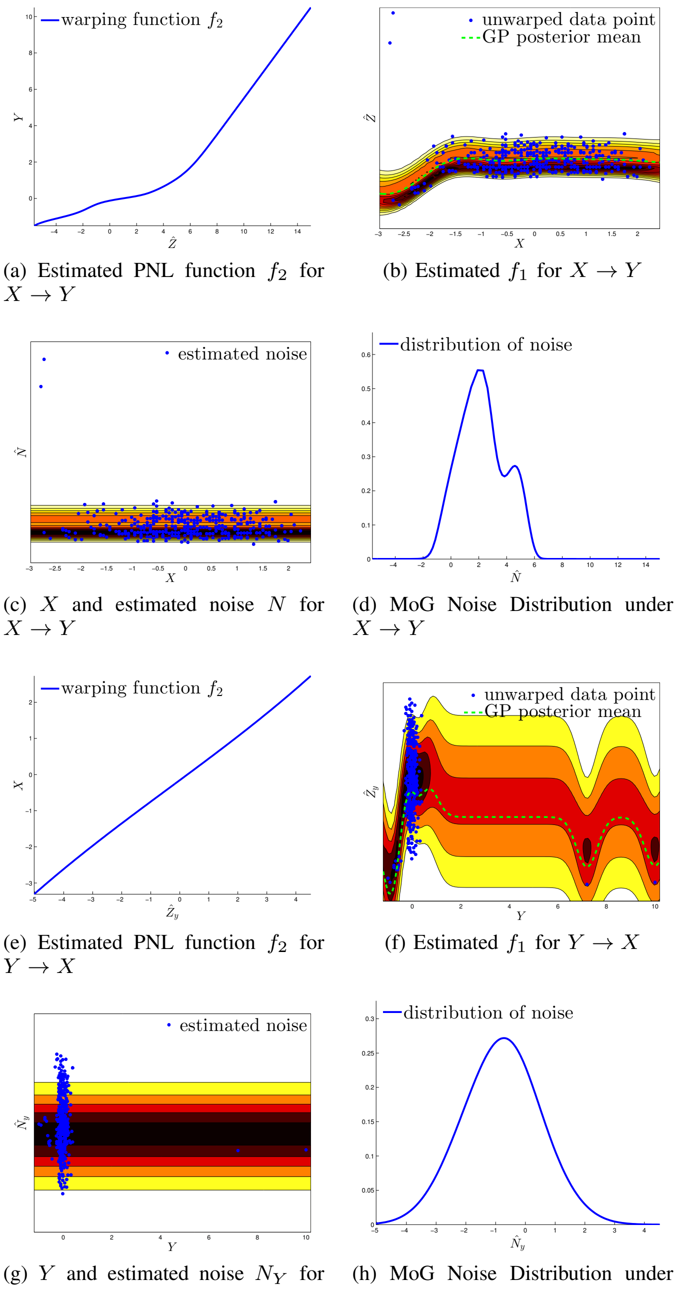
(d) X and estimated noise N for $X \to Y$

(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. 4: 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 one would believe that $X \to Y$. (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.



(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. 5: 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 $Y \to X$.