8

NONLINEAR FUNCTIONAL CAUSAL MODELS FOR DISTINGUISHING CAUSE FROM EFFECT

Kun Zhang

Max Planck Institute for Intelligent Systems, Tübingen, Germany and Carnegie Mellon University, Pittsburgh, PA, USA

Aapo Hyvärinen

Department of Computer Science, University of Helsinki, Helsinki, Finland

8.1 INTRODUCTION

Finding causal directions is a fundamental problem in scientific data analysis and other fields. In general, finding causal directions is extremely complex, but we can make progress by assuming that the causal relationships can only take some special forms.

For simplicity, let us assume that we have only two observed random variables, x and y, where either x is causing y or y is causing x. In particular, we exclude the possibility that there is some kind of bidirectional influence (feedback) between the two; we also exclude the case where both are actually caused by some further, unobserved variable (confounder).

Let us start by considering the very simplest case, where the relationship is assumed *linear*. We thus need to choose between the following two models. The first model assumes that x causes y, and is given by

$$y = \rho x + n \tag{8.1}$$

Statistics and Causality: Methods for Applied Empirical Research, First Edition.
Wolfgang Wiedermann and Alexander von Eye.

while the second model assumes that y causes x and is given by

$$x = \rho y + \tilde{n} \tag{8.2}$$

In both models, the disturbances (also called external influences or noise), denoted by n or $\tilde{n}$, are assumed independent of the regressors x and y, respectively. Without restriction of generality, we can assume that x and y are standardized to zero mean and unit variance. The parameter ρ is then the same in the two models because it is equal to the correlation coefficient.

Choosing between these two models, that is, identifying the causal direction, is a well-known problem that is widely encountered in statistics and machine learning. The problem is usually considered very difficult and perhaps unsolvable, since most analysis assumes that the variables x and y are *Gaussian*, which also implies that the disturbances are Gaussian. Under the Gaussian assumption, the two models are completely symmetric in the sense that the variance explained is equal for the two models, and further, the likelihood is the same for both models (both quantities being simple functions of ρ).

The symmetry between the two models is illustrated in Figure 8.1a, where the points were generated by $y = 2x + n$ and x and n follow the standard Gaussian distribution. We see that the Gaussian data cloud generated by any of the two models looks just the same, which underlines the unidentifiability of the causal direction.

The inability to decide between these two models under the assumptions of linearity and Gaussianity is one of the motivations for the well-known saying that "correlations does not equal causality." However, it is possible to change the situation by modifying any of these two assumptions.

For example, we can make the causal direction identifiable by assuming at least one of the variables (the regressor or the disturbance) in the true model is non-Gaussian. This leads to the theory whose main model is the linear non-Gaussian acyclic model (Shimizu *et al.*, 2006), treated in another chapter of this volume. It is worth noting that there have been several pieces of work in statistics about the asymmetry between two variables in the linear non-Gaussian case. Dating back to 2000, Dodge and Rousson (2000, 2001) considered identifying the correct bivariate linear regression model under the assumption of a non-Gaussian true predictor, which is also addressed in a related chapter of this volume (Dodge and Rousson, 2016). The case of a normal predictor and a non-Gaussian disturbance has been discussed by Wiedermann and Hagmann (2015).

Here, we merely show how the symmetry between the two models is broken, as is illustrated in Figure 8.1b, where the data were generated by $y = 2x + n$ and x and n were obtained by taking the square of standard Gaussian random samples and keeping their original sign. We see that scatterplots for the two models are quite different from each other (note the thin "arms" along either the vertical or the horizontal axis), which gives hope that the causal direction could be identifiable.

Another modification that makes the causal direction identifiable is to assume a nonlinear relationship, which is the topic of this chapter. We start by defining the basic nonlinear model, show how it can be estimated, and then go to a more general theory with more complex nonlinear relationships.

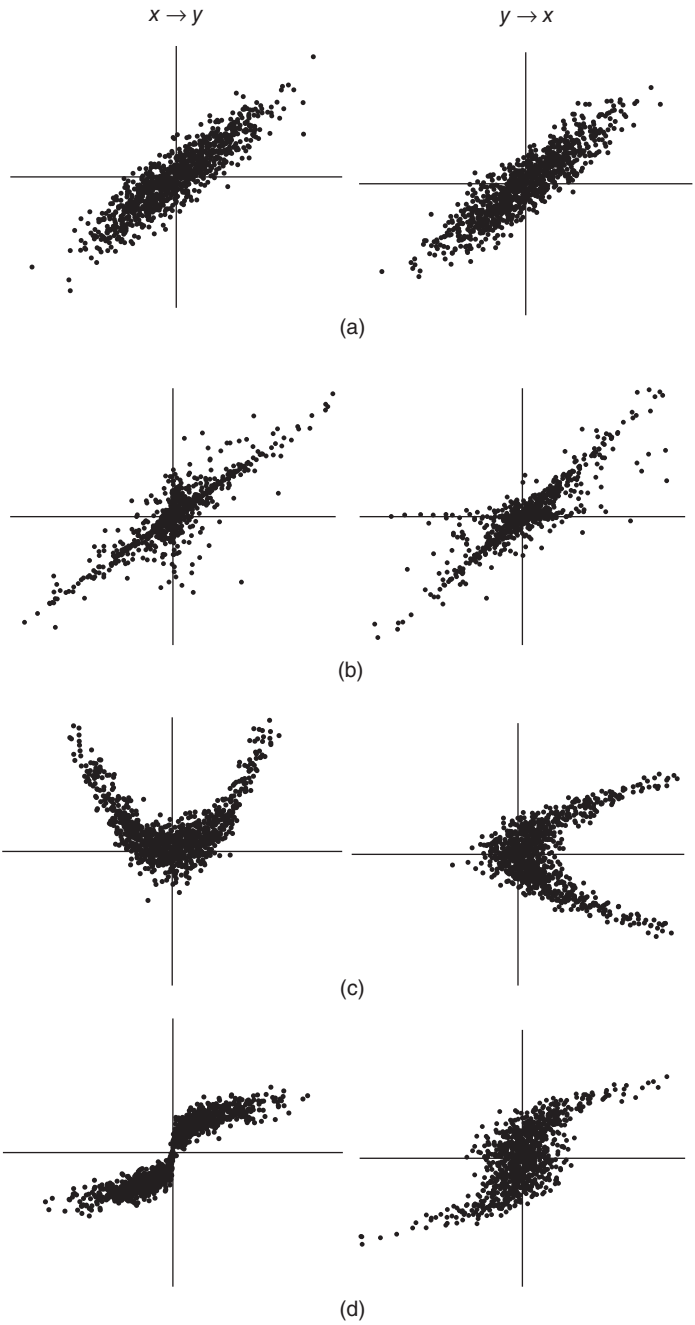

Figure 8.1 Illustration of the effect of the assumptions of linearity and Gaussianity on the identifiability. On the left, we have data generated for the causal direction $x \to y$, and on the right, data generated for the causal direction $y \to x$. The rows correspond to different models: (a) linear Gaussian model, (b) linear non-Gaussian model, (c) the nonlinear model, with two squaring nonlinearity in both directions, (d) the nonlinear model, with cubic root nonlinearity and cubic nonlinearity.

8.2 NONLINEAR ADDITIVE NOISE MODEL

8.2.1 Definition of Model

We start with a particularly simple form of nonlinear relationship, where a scalar-valued nonlinear function, f or g, is taken of the regressor, and the disturbance is added in an additive manner. Thus, we obtain the following two models to choose from. The first model, which we denote by $x \to y$, assumes that x causes y and is given by

$$y = f(x) + n \tag{8.3}$$

while the second model that assumes that y causes x, which we denote by $y \to x$, and is given by

$$x = g(y) + \tilde{n} \tag{8.4}$$

Again, in both models, the disturbances n and $\tilde{n}$, are independent of the regressors x and y, respectively. Here, unlike in the linear case, the functions f and g are likely to be very different from each other. Furthermore, no assumption is made on the distributions of the residuals.

To start with a graphical illustration, see Figure 8.1c. Here, on the left side, we have generated data from $x \to y$ with $f(x) = x^2$ and used Gaussian noise n. We use exactly the same nonlinearity and noise on the right in the direction $y \to x$. This is to illustrate the intuitively quite obvious idea that if the nonlinearity is not invertible (such as squaring), the model is intuitively easy to choose since in the wrong direction, it is not at all of the desired form: nonlinear transform plus independent noise. In fact, it is implausible that y on the right-hand side (generating direction $y \to x$) could have been obtained by adding independent noise on some function of x since any function of x cannot predict y well; in fact here the function of x that predicts y best is $y = 0$. Thus, only the true generating model is at all plausible.

In Figure 8.1d, we have a less drastic, and invertible, nonlinearity (third power) for $x \to y$. Now, since f and g need not be the same in general, a more realistic illustration would take $g = f^{-1}$, which we do here. In particular, we take $f(x) = x^3$ and $g(y) = |y|^{1/3}\text{sign}(y)$. The breaking of symmetry between x and y is seen in the fact that the data distributions are slightly different for the two models, even though they attempted to create the same kind of joint distribution. (We also tried to match the noise levels to make the distributions as similar as possible.) In particular, the noise is seen to "fatten" the regression curve either vertically or horizontally, depending on the direction of causal influence.

8.2.2 Likelihood Ratio for Nonlinear Additive Models

An attractive way of deciding between the two models in Equations (8.1) and (8.2) is to compute their likelihoods and compare them in terms of their ratio. Essentially, we choose the model that has the larger likelihood.

The log-likelihood of the model $x \to y$, for a single data point, can be obtained as the sum of the log-prior of the variable x and the log-likelihood of the residual:

$$\log p(x, y) = \log p_x(x) + \log p_n(y - f(x)) \tag{8.5}$$

Here, it is crucial that we have some kinds of estimates of the log-probability density functions (log-pdf's) of the regressor and the disturbance. Since it is usually more convenient to operate with standardized quantities, so let us denote the log-pdf of the standardized residual by G_n and the log-pdf of x by G_x. Then, the log-likelihood for a single data point can be written as

$$\log p(x, y) = G_x(x) + G_n\left(\frac{y - f(x)}{\sigma_n}\right) - \log \sigma_n \tag{8.6}$$

where we denote the variance of the disturbance by σ_n^2. Choosing the standardized log-pdf's G_x, G_n could be done by modeling the relevant log-pdf's by parametric (Karvanen and Koivunen, 2002) or nonparametric (Pham and Garat, 1997) methods. It is also possible that we have enough prior information on them, so we can fix the G in advance.

Consider a sample $(x_1, y_1), \cdots, (x_T, y_T)$ of data. Let us add together the log-likelihoods of the data points and take the difference, and we obtain the logarithm of the likelihood ratio for the sample as

$$R = \frac{1}{T}\sum_t \left[G_x(x_t) + G_n\left(\frac{y_t - f(x_t)}{\sigma_n}\right) - G_y(y_t) - G_{\tilde{n}}\left(\frac{x_t - g(y_t)}{\sigma_{\tilde{n}}}\right)\right] - \log \sigma_n + \log \sigma_{\tilde{n}} \tag{8.7}$$

where we also need the standardized log-pdf's of y and the disturbance $\tilde{n}$.

The likelihood ratio further depends on the estimated nonlinearities f, g. The estimation of f and g can be done with classic least-squares estimation methods, fitting some nonlinear (nonparametric) regression model on the sample. Such regression methods are independent of any developments in this chapter. A large number of nonparametric methods have been developed in the literature; see Hoyer *et al.* (2009) for an example.

8.2.3 Information-Theoretic Interpretation

The likelihood ratio has a simple information-theoretic interpretation, which also means we can use well-known information-theoretic approximations for its practical computation in the case where we do not want to postulate functional forms for the G's.

In fact, if we take the asymptotic limit of the likelihood ratio ($T \to \infty$), we obtain asymptotically

$$R \to -H(x) - H\left(\frac{n}{\sigma_n}\right) + H(y) + H\left(\frac{\tilde{n}}{\sigma_{\tilde{n}}}\right) - \log \sigma_n + \log \sigma_{\tilde{n}} \tag{8.8}$$

where we denote differential entropy by H. The differential entropy, defined as $H(x) = -\int p(x) \log p(x) dx$, is the fundamental information-theoretic quantity for continuous-valued variables.

We can go back to the nonstandardized quantities (which can here be done by using the fundamental transformation formula $H(\alpha x) = H(x) + \log \alpha$, and we further obtain a very simple equivalent expression:

$$R \to -H(x) - H(n) + H(y) + H(\tilde{n}) \tag{8.9}$$

Here, we see that determining causal direction is related to finding the direction that corresponds to minimum entropy in the sense of the sum of the marginal entropies of the regressor and the disturbance; we have more on this connection next.

The practical utility of this connection is that we can approximate the likelihood ratio using any general, possibly nonparametric, approximations of differential entropy. Several such approximations have been developed; for example, we can use the maximum entropy approximations by Hyvärinen (1998), which are computationally simple. In fact, we only need to approximate one-dimensional differential entropies, which is much simpler than approximating two-dimensional entropies.

This information-theoretic formulation also leads to a simple intuitive interpretation of the likelihood ratio. It is well known that in the space of probability distributions of unit variance, differential entropy is maximized by Gaussian distribution. This is why (negative) differential entropy is often used as a measure of non-Gaussianity. In our case, we can thus interpret the asymptotic limit of the log-likelihood ratio in terms of non-Gaussianities and errors in regression:

$$\begin{aligned} &R \to \\ &\text{nongaussianity}(x) + \text{nongaussianity}(\text{residual in } x \to y) - \log(\text{error in } x \to y) \\ &- [\text{nongaussianity}(y) + \text{nongaussianity}(\text{residual in } y \to x) - \log(\text{error in } y \to x)] \end{aligned}$$

Intuitively, this means that

(1) if the non-Gaussianities are negligible, we choose the direction in which the error in the regression is smaller;
(2) if the errors in the regression are almost equal, we choose the direction of causality in which the sum of non-Gaussianities of the regressor and residual is maximized;
(3) in the general case, we have a sum of these two criteria: error in regression and non-Gaussianity.

An interesting point here is that in the linear non-Gaussian case, the errors in the regression are always equal, and thus, choosing the direction is solely based on maximizing the non-Gaussianity. In contrast, in the nonlinear case, the errors in the regressions can be the decisive factor in the identification. This is intuitively clear in Figure 8.1c, where the right direction leads to a small regression error, while in the wrong direction, the regression is catastrophically bad.

8.2.4 Likelihood Ratio and Independence-Based Methods

An alternative approach to nonlinear additive models is provided by the independence-based method by Hoyer *et al.* (2009). In such methods, the idea is to use the independence of the regressor and the disturbance in each model as the selection criterion: the model in which the residual (i.e., estimate of disturbance) is more independent of the regressor is chosen (again, assuming that some nonparametric regression method is used to estimate f and g; see Section 8.3.3.)

The fundamental information-theoretic quantity for measuring the independence of two random variables is mutual information, defined as

$$I(u, v) = H(u) + H(v) - H(u, v) \tag{8.10}$$

which is always nonnegative and zero if and only if the two variables are independent. Here, $H(u, v)$ is the joint entropy, which is simply the entropy of the random vector consisting of (u, v).

In fact, the likelihood ratio can be interpreted from the viewpoint of such maximization of independence. Using basic information-theoretic properties, we have under $x \to y$

$$\begin{aligned} H(x, y) &= H(x) + H(y|x) = H(x) + H(y - f(x)|x) \\ &= H(x) + H(n|x) = H(x, n) \end{aligned} \tag{8.11}$$

and by symmetry, this is equal to $H(y, \tilde{n})$. Now if we can consider the difference between the mutual information of the regressors and residuals in the two directions and obtain

$$\begin{aligned} I(x, n) - I(y, e) &= H(x) + H(n) - H(x, n) - [H(y) + H(e) - H(y, e)] \\ &= H(x) + H(n) - H(y) - H(e) \\ &= H(x) + H\left(\frac{n}{\sigma_n}\right) - H(y) - H\left(\frac{\tilde{n}}{\sigma_{\tilde{n}}}\right) \\ &\quad + \log \sigma_n - \log \sigma_{\tilde{n}} \end{aligned} \tag{8.12}$$

where two terms equal to $H(x, y)$ cancel. Here, we see that asymptotically, the objective derived from the likelihood ratio is equal to the difference of the two mutual information (with sign reversed). Its sign tells which mutual information is larger and, in particular, in which direction the residual of the regression is more independent. Thus, using the likelihood ratio is equivalent to using mutual information as independence measure in the methods by Hoyer *et al.* (2009). We will elaborate more on this in Section 8.4.

The aforementioned developments thus show that when comparing independencies of the residuals such as Hoyer *et al.* (2009), it is not necessary to explicitly estimate mutual information; estimation of one-dimensional entropies leads to an

equivalent result. This is very important from a practical viewpoint, since estimating one-dimensional entropies is much easier than estimating two-dimensional quantities such as mutual information.

8.3 POST-NONLINEAR CAUSAL MODEL

Obviously, it is important to use sufficiently general functional models in causal discovery: if the assumed functional causal model is too restrictive to properly approximate the true data-generating process, the causal discovery results may be misleading. Therefore, if specific knowledge about the data-generating mechanism is not available, we should attempt to fit a model that is as general as possible. Post-nonlinear (PNL) causal models offer an interesting generalization of the nonlinear additive model of the previous section.

8.3.1 The Model

The PNL causal model consists of a nonlinear influence from the cause, a noise or disturbance, and–in contrast to the model above–a possible sensor or measurement distortion in the observed variables (Zhang and Hyvärinen, 2009b, 2010). The effect y is generated by a post-nonlinear transformation of the nonlinear effect of the cause x with additive noise n:

$$y = f_2(f_1(x) + n) \tag{8.13}$$

where both f_1 and f_2 are nonlinear functions and f_2 is assumed to be invertible. The post-nonlinear transformation f_2 represents the sensor or measurement distortion, which is frequently encountered in practice. A slightly more restricted version of the model, in which the inner function, f_1, is also assumed to be invertible, was proposed in Zhang and Chan (2006) and applied to causal analysis of stock returns.[1]

The PNL causal model has the most general form among all well-defined functional causal models in which the causal direction has been shown to be identifiable under mild assumptions. Clearly, it contains the linear model and the nonlinear additive noise model as special cases. The multiplicative noise model, $y = x \cdot n$, where all involved variables are positive, is another special case: the multiplicative noise model can be written as $y = \exp(\log x + \log n)$, where $\log n$ is considered as a new noise term, $f_1(x) = \log(x)$, and $f_2(\cdot) = \exp(\cdot)$.

Next, we discuss the identifiability conditions of the causal direction for the PNL causal model, which naturally contain those for the linear model and nonlinear additive noise model as special cases.

[1] In Zhang and Chan (2006) both functions f_1 and f_2 are assumed to be invertible; this causal model, as a consequence, can be estimated by making use of post-nonlinear independent component analysis (PNL-ICA; Taleb and Jutten, 1999), which assumes that the observed data are componentwise invertible transformations of linear mixtures of the independence sources to be recovered.

8.3.2 Identifiability of Causal Direction

The identifiability conditions of the causal direction according to the PNL causal model were established by a proof by contradiction (Zhang and Hyvärinen, 2009b). We assume the causal model holds in both directions $x \to y$ and $y \to x$ and show that this implies some very strong conditions on the distributions and functions involved in the model. Therefore, if the data are generated according to the PNL causal model in settings not fulfilling those strong conditions, the backward direction does not follow the model, and the causal direction can be determined. We will next explain this in more detail.

Assume that the data (x, y) is generated by the PNL causal model with the causal relation $x \to y$ in (8.13). Moreover, let us assume (by contradiction) that the backward direction, $y \to x$, also follows the PNL causal model with independent noise. That is,

$$x = g_2(g_1(y) + \tilde{n}) \tag{8.14}$$

where y and $\tilde{n}$ are independent, g_1 is nonconstant, and g_2 is invertible.

Equations (8.13) and (8.14) define the transformation from (x, n) to $(y, \tilde{n})$; as a consequence, using the change-of-variable technique, $p_{y,\tilde{n}}$ can be expressed in terms of $p_{x,n} = p_x p_n$. The identifiability results were derived by making use of linear separability of the logarithm of the joint density of independent variables, that is, for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density is diagonal everywhere (Lin, 1998). Since y and $\tilde{n}$ are assumed to be independent, $\log p_{y,\tilde{n}}$ then follows such a linear separability property. This implies that the second-order partial derivative of $\log p_{y,\tilde{n}}$ w.r.t. y and $\tilde{n}$ is zero. It then reduces to a differential equation of a bilinear form. Under certain technical assumptions (e.g., p_n is positive on $(-\infty, +\infty)$), the solution to the differential equation gives all cases in which the causal direction is *not* identifiable according to the PNL causal model. Table 1 in Zhang and Hyvärinen (2009b) summarizes all five nonidentifiable cases. The first one is the linear-Gaussian case, in which the causal direction is well known to be nonidentifiable. Roughly speaking, to make one of those cases true, one has to adjust the data distribution and the involved nonlinear functions very carefully.

In other words, in the generic case, the causal direction is identifiable if the data were generated according to the PNL causal model. Simulations results were further presented in Zhang and Hyvärinen (2009b) to verify the established identifiability results.

8.3.3 Determination of Causal Direction Based on the PNL Causal Model

The commonly used approach to distinguishing cause from effect with nonlinear functional causal models consists of two steps, which are similar for both the nonlinear additive noise model and post-nonlinear model. First, one fits the nonlinear regression model on the data for both hypothetical causal directions, obtaining estimates for f and g. The second step is to do a statistical analysis of the regressors and the residuals to determine the causal direction.

For the nonlinear additive noise model, the functions f and g are usually estimated by performing quite conventional Gaussian process (GP) regression (Hoyer *et al.*, 2009). (For details on GP regression, one may refer to Rasmussen and Williams, 2006) In contrast, estimation of the PNL causal model (8.13) has several indeterminacies: the sign, mean, and scale of the noise term, and accordingly, the sign, location, and scale of f, g are arbitrary. In the estimation procedure, one may impose certain constraints to avoid such indeterminacies in the estimate. However, we should note that in principle, we do not care about those indeterminacies in the causal discovery context, since they do not change the statistical independence or dependence property between the estimated noise and the hypothetical cause.

It is well known that for linear regression, the maximum likelihood estimator of the coefficient is still statistically consistent even if the noise distribution is erroneously assumed to be Gaussian. However, this may not be the case for general nonlinear models. As shown in (Zhang *et al.* 2016, Section 3.2), if the noise distribution is misspecified, the estimated PNL causal model (8.13) may not be statistically consistent, even when the indeterminacies in the estimate discussed earlier are properly tackled. Therefore, the noise distribution should be adaptively estimated from data, if the true one is not known *a priori*.

Regarding the statistical analysis of the regressor and the residuals, performing independence tests between the estimated noise and hypothetical cause is one approach (Hoyer *et al.*, 2009), (Zhang and Hyvärinen, 2009b). A commonly used option is the Hilbert-Schmidt information criterion (HSIC; (Gretton *et al.*, 2005)), although many others could be used. In fact, for nonlinear additive noise models, as we discussed in Section 8.2.2, following Hyvärinen and Smith (2013), the independence can be evaluated using one-dimensional entropy estimators as well.

Considering concrete implementations in the literature, Zhang and Hyvärinen (2009b) proposed to estimate the PNL causal model (8.13) by mutual information minimization with the involved nonlinear functions represented by multilayer perceptrons (MLPs). Later, Zhang *et al.* (2016) proposed to estimate the PNL causal model by making use of warped Gaussian processes with a flexible noise distribution, which is represented by a mixture of Gaussians. We call these two implementations PNL-MLP and PNL-WGP-MoG, respectively.

8.4 ON THE RELATIONSHIPS BETWEEN DIFFERENT PRINCIPLES FOR MODEL ESTIMATION

So far, we have mainly discussed the identifiability of the causal direction in the two-variable case, and it should be noted that the results can be readily extended to the case with an arbitrary number of variables, as shown in Peters *et al.* (2011). The basic idea is that no matter how many variables are involved in the system, when we are interested in a particular pair of directly connected variables, it becomes the two-variable case if the values of relevant variables are fixed.

Maximum likelihood is usually used to fit the functional causal model together with a directed acyclic graph (DAG) to the given data. Not surprisingly, the negative

likelihood (with the distribution of the noise adaptively estimated from data) is equivalent to the mutual information between the estimated noise terms, as stated in Theorem 3 in Zhang *et al.* (2016). The higher the likelihood, the less dependent the estimated noise terms. (Note that the root variables in the DAG are also counted as noise terms.)

On the other hand, traditionally, it has been noted that under the causal Markov condition, which states that each variable is independent from its nondescendants conditioning on its parents, and the faithfulness assumption, one could recover an equivalence class of the underlying causal structure based on conditional independence relationships of the variables (Spirtes *et al.*, 2001, Pearl, 2000). This is known as the constraint-based approach to causal discovery. How are these principles, including mutual independence of the estimated noise terms and the causal Markov condition, related to each other? Next, we will answer this question, and the results in this section hold for an arbitrary number of variables.

In the following, we consider optimization over different DAG structures to find the causal structure. We assume we have infinite data, and we optimally fit the nonlinear functions f_i according to the DAG structure given, using some hypothetical method, which is statistically consistent. Then the question is how the statistical properties of the estimated noise terms (residuals) are related to the conditional independence properties of the variables x_i, for each particular DAG.

Suppose (first) that we fit the nonlinear additive noise model given the DAG structure, that is,

$$x_i = f_i(pa_i) + n_i \tag{8.15}$$

where pa_i represents the direct causes of x_i, to the data, that is, parents in the DAG. It has been shown that mutual independence of the estimated residuals and conditional independence between observed variables (together with the independence between n_i and pa_i) are equivalent; furthermore, they are achieved if and only if the total entropy of the disturbances is minimized (Zhang and Hyvärinen, 2009a). More specifically, when fitting the model (8.15) with a hypothetical DAG causal structure to the given variables $x_1, \cdots, x_K$, the following three properties are equivalent:

(i) The estimated noise terms n_i are mutually independent.

(ii) The total entropy of the estimated noise terms, that is, $\sum_i H(n_i)$, is minimized, with the minimum being equal to $H(x_1, \cdots, x_K)$.

(iii) The causal Markov condition holds (i.e., each variable is independent of its nondescendants in the DAG conditioning on its parents), and in addition, the noise term in x_i is independent of the parents of x_i.

Let us then consider the PNL causal model. When one fits the PNL causal model

$$x_i = f_{i2}(f_{i1}(pa_i) + n_i) \tag{8.16}$$

to the data, the scale of the noise terms as well as f_{i1} is arbitrary, since f_{i2} is also to be estimated. Consequently, unlike for the nonlinear additive noise model, in the PNL

causal model context, it is not meaningful to talk about the total entropy of the noise terms (see condition (iii)). However, as shown in Zhang and Hyvärinen (2009b), when fitting the PNL causal model with a hypothetical DAG causal structure to the data, we still have the equivalence between conditions (i) and (iii).

The next question is how to estimate a functional causal model for more than two variables in practice. one approach 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 superexponentially along with the number of variables. Smarter approaches are thus needed.

The aforementioned theorem suggests a simpler two-step method to find the causal structure implied by the PNL causal model. We use here the relationship between mutual independence of the noise terms and the causal Markov condition combined with the independence between each noise term and its associated parents. One first uses the constraint-based approach (Spirtes *et al.*, 2001), (Pearl, 2000) to find the Markov equivalent class from conditional independence relationships given by some nonparametric conditional independence tests; for instance, one can adopt the kernel-based test (Zhang *et al.*, 2011). This approach first finds the skeleton of the causal graph by removing the edge between a pair of variables, if there exists some subset of variables (including the empty set) given that they are conditionally independent. It then uses the orientation rules to find the causal directions of some edges. The PNL causal model is then used to identify the causal directions that cannot be determined in the first step: for each DAG contained in the equivalent class, we estimate the noise terms and determine whether this causal structure is plausible by examining whether the disturbance in each variable x_i is independent of the parents of x_i. Consequently, one avoids the exhaustive search over all possible causal structures and high-dimensional statistical tests of mutual independence of all noise terms.

8.5 REMARK ON GENERAL NONLINEAR CAUSAL MODELS

We have discussed several functional causal models, namely, the linear model, nonlinear additive noise model, and PNL causal model. Now let us discuss the possibility of doing causal discovery with the general form of functional causal models. A functional causal model represents the effect y as a function of the direct causes x and some unmeasurable noise (Pearl, 2000):

$$y = f(x, n; \theta_1) \tag{8.17}$$

where n is the noise that is assumed to be independent of x, the function $f \in \mathcal{F}$ explains how y is generated from x, $\mathcal{F}$ is an appropriately constrained functional class, and θ_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.

In the functional causal model (8.17), the noise term is assumed to be independent of the cause. If for the reverse direction, one cannot find a noise term that is independent of the hypothetical cause (which is y), then we can determine the true causal direction or distinguish cause from effect. As discussed earlier, in general, this is the case for the PNL causal model, as well as for the linear and nonlinear models with additive noise. Unfortunately, this is not the case if we do not impose any constraint on the function f.

As discussed in Hyvärinen and Pajunen (1999), given *any* two random variables x and y with continuous support, no matter how they are related, one can always construct another variable, denoted by $\mathring{n}$, which is statistically independent of x. In Zhang *et al.* (2016), the class of functions to produce such an independent variable $\mathring{n}$ (or called independent noise term in our causal discovery context) was given, and it was shown that this procedure is invertible: y is a function of x and $\mathring{n}$.

This is also the case for the hypothetical causal direction $y \to x$: we can also always represent x as a function of y and an independent noise term, if the functional form is not properly constrained. 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 hypothetical cause holds for only one direction. Examples of such constraints include the linear model, the nonlinear additive noise model, and the PNL causal model discussed earlier. As we have already seen, under appropriate assumptions, the constraints of additive noise and the PNL data-generating model serve such a goal.

8.6 SOME EMPIRICAL RESULTS

Various nonlinear functional causal models have been used to distinguish cause from effect on the cause-effect pairs available at `http://webdav.tuebingen.mpg.de/cause-effect/`. They consist of different data pairs, for which the causal direction is believed to be known, for testing the performance of causal detection algorithms. They are from different scientific disciplines including climate analysis, finance, and computer science. Here let us summarize the results reported in Zhang *et al.* (2016). The exploited 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 Gaussian processes with Gaussian noise, denoted by PNL-WGP-Gaussian, the PNL causal model estimated by warped Gaussian 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) is reported as follows:

PNL-MLP: 70%
PNL-WGP-Gaussian: 67%
PNL-WGP-MoG: 76%
ANM: 63%
GPI: 72%
IGCI: 73%

One can see that all results are better than chance, illustrating the effectivity of using functional causal models to distinguish cause from effect. Here PNL-WGP-MoG gives the best performance among these methods.

8.7 DISCUSSION AND CONCLUSION

We have given a survey of functional causal models that enable us to fully identify the causal structure from observational data. We focused on the two-variable case, where the task is to distinguish cause from effect. We have reviewed the linear non-Gaussian causal model, nonlinear additive noise model, and the post-nonlinear causal model, listed from the most to the least restrictive. We addressed the identifiability of the causal direction: for those three models, in the generic case, the backward direction does not admit an independent noise term, and as a consequence, it is possible to distinguish cause from effect. We have also briefly discussed the procedure to achieve so, which consists of fitting the functional model and performing an independence test between the estimated noise and the hypothetical cause. For nonlinear additive noise models, we have also presented a likelihood-ratio-based approach to determining the causal direction.

There are some open problems along this line of research. First, one can consider functional causal models as a way to represent the conditional distribution of the effect, given the cause. Can we then find hints as to the causal direction directly from the data distribution? Or, in other words, can we find a general way to characterize the causal asymmetry directly in terms of certain properties of the data distribution? An attempt to do so is to make use of the so-called exogeneity property of a causally sufficient causal system (Zhang *et al.*, 2015).

Secondly, note that nonlinear functional causal models are usually intransitive. That is, if both causal processes $x_1 \to x_2$ and $x_2 \to x_3$ admit a particular type of functional causal model, the process $x_1 \to x_3$ does not necessarily follow the same model. (Linear models are transitive.) This could be a potential issue of functional-causal-model-based causal discovery: it may fail to discover indirect

causal relations. (Here, by direct causal relations, we mean the causal relations in which only a single noise variable is involved.) On the other hand, this may be a benefit of using functional causal models for causal discovery, in that it is possible to detect the existence of the causal intermediate variable and further recover it. But how to do so is currently unclear.

In this chapter, we were concerned with causal discovery in the continuous case. In the discrete case, if one knows precisely what model class generated the effect from cause, which, for instance, may be the noisy AND or noisy XOR gate, then under mild conditions, the causal direction can be easily seen from the data distribution. Consider binary variables and take the noisy AND gate as the causal process. Then the probability of the effect variable taking value 1 is smaller than (or equal to, if the noise only takes value 1) that for the cause variable. However, generally speaking, if the precise model class of the causal process is unknown, it is difficult to recover the causal direction from observed data in the discrete case, especially when the cardinality of the variables is small. As an illustration, consider the situation where the causal process first generates continuous data and discretizes such data to produce the observed discrete ones. It is then not surprising that certain properties of the causal process are lost due to discretization, making causal discovery more difficult.

Finally, developing efficient methods for causal discovery of more than two variables based on functional causal models is an important step toward solving large-scale real-world causal analysis problems in various domains including neuroscience and biology. To make causal discovery computationally efficient, one may have to limit the complexity of the causal structure, say, limit the number of direct causes of each variable. Even so, a smart optimization procedure instead of exhaustive search is still missing in the literature.

The package for estimating the post-nonlinear causal model and causal direction identification based on this model is available at `http://webdav.tuebingen.mpg.de/causality/CauseOrEffect_NICA.rar` (with nonlinear functions represented by MLPs) or `http://people.tuebingen.mpg.de/kzhang/warpedGP.zip` (estimated by warped Gaussian processes with mixture-of-Gaussian noise).

REFERENCES

Dodge, Y. and Rousson, V. (2000) Direction dependence in a regression line. *Communications in Statistics: Theory and Methods*, **29**, 1957–1972.

Dodge, Y. and Rousson, V. (2001) On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician*, **55**, 51–54.

Dodge, Y. and Rousson, V. (2016) Recent developments on the direction of a regression line, in *Statistics and Causality: Methods for Applied Empirical Research* (eds W. Wiedermann and A. von Eye), John Wiley & Sons, Inc.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005) Measuring statistical dependence with Hilbert-Schmidt norms, in *Algorithmic Learning Theory: 16th International Conference* (eds S. Jain, H. Simon, and E. Tomita), Springer-Verlag, Berlin, pp. 63–78.

Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J., and Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models, in *Advances in Neural Information Processing Systems 21*, Vancouver, BC, Canada, pp. 689–696.

Hyvärinen, A. (1998) New approximations of differential entropy for independent component analysis and projection pursuit, in *Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge, MA, pp. 273–279.

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

Hyvärinen, A. and Smith, S. (2013) Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, **14**, 111–152.

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

Karvanen, J. and Koivunen, V. (2002) Blind separation methods based on Pearson system and its extensions. *Signal Processing*, **82**, 663–573.

Lin, J. (1998) Factorizing multivariate function classes, in *Advances in Neural Information Processing Systems 10*, MIT Press, Cambridge, MA, pp. 563–569.

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

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

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2011) Identifiability of causal graphs using functional models, in Proceedings of UAI 2011, pp. 589–598.

Pham, D. and Garat, P. (1997) Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, **45** (7), 1712–1725.

Rasmussen, C. and Williams, C. (2006) *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA.

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

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

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

Wiedermann, W. and Hagmann, M. (2015) Asymmetric properties of the Pearson correlation coefficient: correlation as the negative association between linear regression residuals. *Communications in Statistics: Theory and Methods*, in press.

Zhang, K. and Chan, L. (2006) Extensions of ICA for causality discovery in the Hong Kong stock market, in Proceedings of the 13th International Conference on Neural Information Processing (ICONIP 2006).

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

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

Zhang, K. and Hyvärinen, A. (2010) Distinguishing causes from effects using nonlinear acyclic causal models, in JMLR Workshop and Conference Proceedings, vol. 6, pp. 157–164.

Zhang, K., Peters, J., and Janzing, D., and Schölkopf, B. (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.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (2016) On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. **7** (2), ACM, New York.

Zhang, K., Zhang, J., and Schölkopf, B. (2015) Distinguishing cause from effect based on exogeneity, in Proceedings of the 15th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2015).