Learning Bivariate Functional Causal Models

Olivier Goudet, Diviyan Kalainathan, Michèle Sebag, Isabelle Guyon

Chapter 3
Learning Bivariate Functional Causal Models

Olivier Goudet, Diviyan Kalainathan, Michèle Sebag, and Isabelle Guyon

3.1 Introduction

A natural approach to address the cause-effect pair problem is from a reverse engineering perspective: given the available measurements $\{(x_i, y_i)\}_{i=1}^n$ of the two variables X and Y, the task is to discover the underlying causal process that led the variables to take the values they have. To find this model, one needs implicitly to answer two questions:

1. First, is there a causal relationship between X and Y and what is the *causal direction*? Was X generated first and then Y generated from X, or the opposite?
2. Second, what is the *causal mechanism* that can explain the behavior of the system? *How* was Y generated from X or X generated from Y?

Therefore this approach for causal discovery goes beyond finding the causal structure, as it requires also to define a generative model of the data, which do not seem mandatory at first sight if one is only interested in finding if $X \rightarrow Y$ or $Y \rightarrow X$.

This type of generative model has notably been formalized with the framework of Functional Causal Models (FCMs) [21], also known as Structural Equation Models (SEMs), that can well represent the underlying data-generating process, supports interventions and allows counterfactual reasoning.

O. Goudet (✉) · D. Kalainathan · M. Sebag
Team TAU - CNRS, INRIA, Université Paris Sud, Université Paris Saclay, Orsay, France
e-mail: Diviyan.kalainathan@lri.fr; sebag@lri.fr

I. Guyon
Team TAU - CNRS, INRIA, Université Paris Sud, Université Paris Saclay, Orsay France

ChaLearn, Berkeley, CA, USA
e-mail: guyon@chalearn.org

3.1.1 Toy Example: Recovering the Underlying Generative Process

Let us first introduce a simple cause-effect example depicted on Fig. 3.1. This bivariate example was generated using an artificial causal mechanism from one variable to another.[1] Interestingly enough, the correlation coefficient between X and Y is zero, but one can immediately see what could be the causal direction.

Indeed it seems very intuitive to prefer the causal direction $X \rightarrow Y$. And this is true in this case as the data were generated from X to Y according the following stochastic process with quadratic deterministic mechanism:

$$X \sim \mathscr{U}(-1, 1) \tag{3.1}$$

$$N_Y \sim \mathscr{U}(-1, 1)/3 \tag{3.2}$$

$$Y := 4 \times (X^2 - 0.5)^2 + N_Y, \tag{3.3}$$

where $\mathscr{U}(-1, 1)$ denotes uniform distribution between -1 and 1. Here we use the symbol ":=" when writing the model to signify that it has to be seen as an assignment from cause to effect.

The correlation coefficient is equal to zero, but there is a nonlinear dependency between X and Y. Moreover, if one assumes that this dependency is due to the influence of one of the variable on the other and not due to a third variable

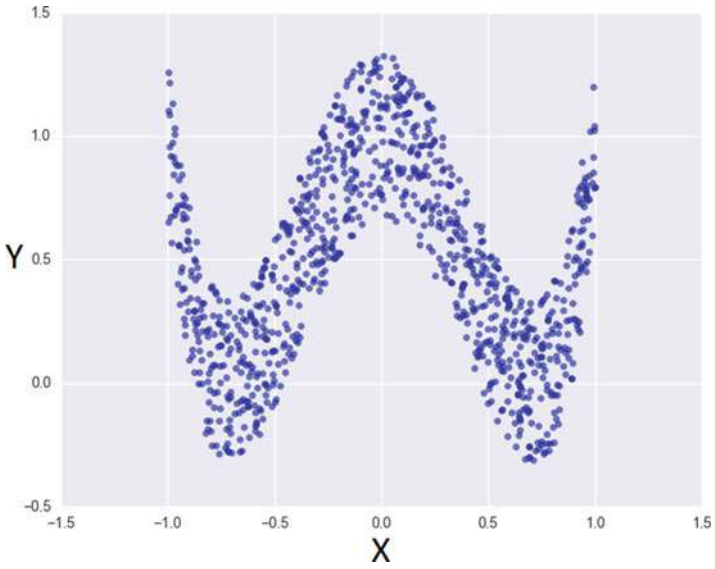

Fig. 3.1 Data generated with the quadratic model of Eq. (3.3)

[1] Example coming from https://en.wikipedia.org/wiki/Correlation_and_dependence.

influencing both, one of the two following causal hypotheses holds as stated in Chap. 1 of this book:

- $Y := f_Y(X, N_Y)$ with $N_Y \perp\!\!\!\perp X$ (hypothesis 1, $X \to Y$),
- $X := f_X(Y, N_X)$ with $N_X \perp\!\!\!\perp Y$ (hypothesis 2, $Y \to X$),

where the noise variable N_Y (respectively N_X) summarizes *all the other unobserved influences* on X (respectively on Y).

In order to recover the causal mechanism, one can propose to search in the class of polynomial functions of degree 4 by fitting regression models $y = \hat{f}_Y(x)$ and $x = \hat{f}_X(y)$ on the data with mean squared error loss (Fig. 3.2).

The expected mean squared error is lower for the model $\hat{f}_Y$ from X to Y than for the other model $\hat{f}_X$ from Y to X. The residuals of each polynomial regression are displayed on Fig. 3.3.

To some extent, as the residual $Y - \hat{f}_Y(X)$ is independent of X, one can build an explanatory model of the data as : $Y := \hat{f}_Y(X) + N_Y$, with $\hat{f}_Y$ quadratic and with almost $N_Y \perp\!\!\!\perp X$ and $N_Y \sim \mathscr{U}(-1, 1)/3$. In this case the underlying data generative process of the data corresponding to the *true* model (see Eq. (3.3)) is recovered. The causal direction $X \to Y$ is identified and one has also built a simulator close to the *true* mechanism (up to small parameter adjustments) that can be used to simulate the effect of interventions on the system. Indeed with this functional model, one can now compute $Y^{do(X=x)} = \hat{f}_Y(x) + N_Y$, with $N_Y \sim \mathscr{U}(-1, 1)/3$. It should give results on average similar to the *true* model of Eq. (3.3).

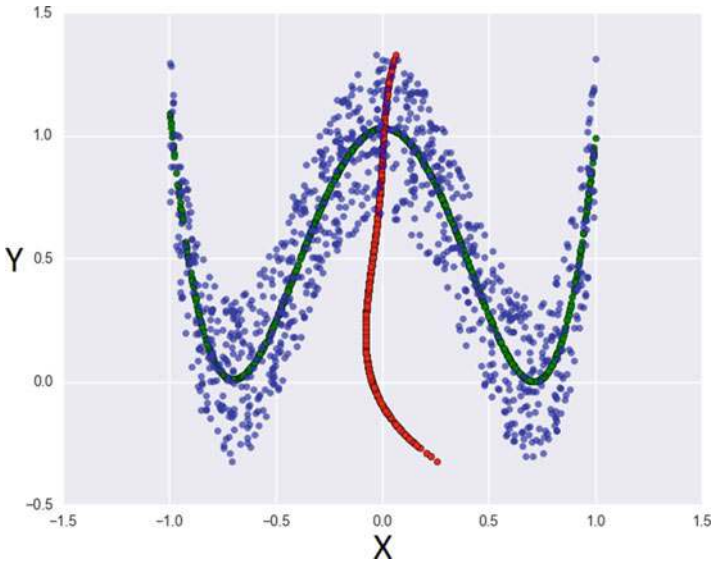

Fig. 3.2 Data generated with the *true* quadratic model (blue points). Polynomial fit of degree 4 of Y on X is depicted as green curve. Polynomial fit of degree 4 of X on Y is depicted as red curve. The *best* fit appears with the green curve. Better seen in color

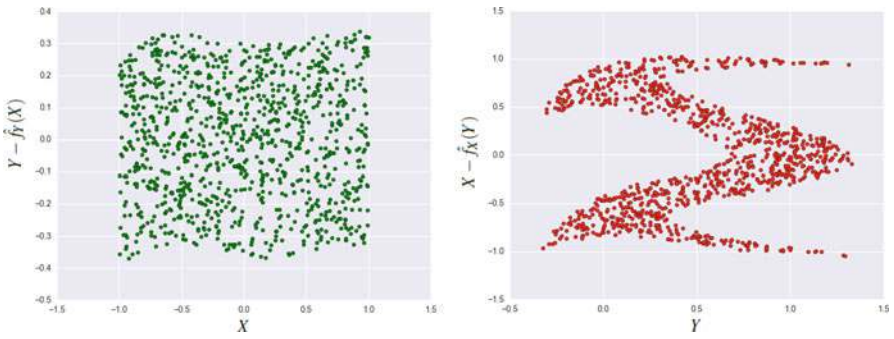

Fig. 3.3 Left: error $Y - \hat{f}_Y(X)$ is almost independent of X. Right: error $X - \hat{f}_X(Y)$ is not independent of Y. Better seen in color

In the opposite direction, as the residual is not independent of Y, one cannot build any model such as $X := \hat{f}_X(Y) + N_X$ with $N_X \perp\!\!\!\perp Y$. However we will see later in this chapter that it is possible to prove that a model $X := \hat{f}_X(Y, N_X)$ with $N_X \perp\!\!\!\perp Y$ exists in this case, but its expression may be more complex as the noise term is not additive and the mechanism $\hat{f}_X$ is not continuous on its definition domain. Therefore, from an explanatory perspective, the additive noise model $Y := \hat{f}_Y(X) + N_Y$ (hypothesis 1) offers a simpler explanation of the phenomenon than the other explanation $X := \hat{f}_X(Y, N_X)$ (hypothesis 2) with a non additive noise. For this reason, it seems intuitive to "prefer" hypothesis 1 to hypothesis 2. It is not a formal proof of causality, as it comes from an "Occam razor principle" that favors a simple explanation with the prior assumption that an additive noise model is simpler than a non additive noise model [12]. Another notion of simplicity could be preferred such as multiplicative noise $Y := f_Y(X) \times N_Y$ as explained later in this chapter.

Given the available measurements, the goal is to recover an *explanatory model* of the data by using statistical tools to test causal hypotheses. However, even if in this toy example the preferred explanatory model also corresponds to the model with the best predictive power, one has to keep in mind that this is not always the case.

3.1.2 Real Example: To Explain or to Predict?

Let us introduce now a real example well known in "the cause-effect pair community": the first cause-effect real pair of the Tübingen database.[2] Figure 3.4 displays collected data on altitude (X-axis) and temperature (Y-axis) in the atmosphere from 349 meteorological stations in Germany over the years 1961–1990.

[2]This database is composed of more than one hundred real cause-effect pairs with known ground truth and is available online at https://webdav.tuebingen.mpg.de/cause-effect/.

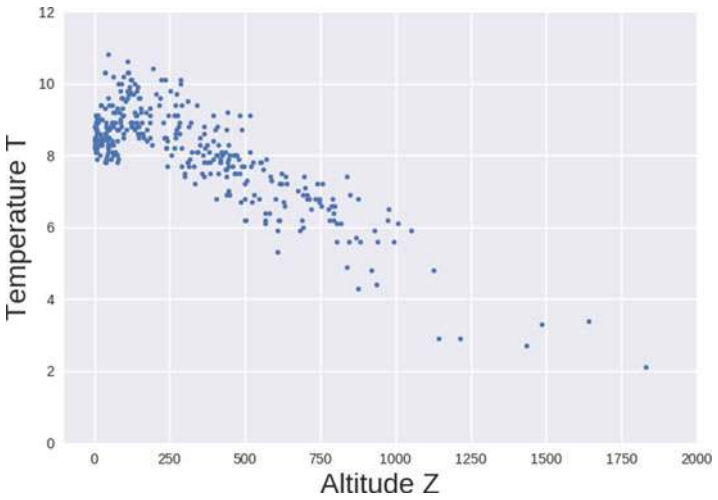

Fig. 3.4 349 real couples of points (altitude, temperature) collected from meteorological stations in Germany over the years 1961–1990

T is the temperature and Z the corresponding altitude. We assume for this simple example that the observed dependency between T and Z is only due to a direct causal relation from one of the two variables to the other. We will see later in this chapter that it is not actually obvious and the model may be actually more complex, as at least one hidden variable latitude may also have an impact on both variables.

In order to test causal hypotheses, the data are re-scaled with zero mean and unit variance, the dataset are split a large number of times between train (80%) and test sets (20%) and two alternative nonlinear Gaussian process regression models are learned on the train set.[3] When regressing the temperature on altitude one obtains the model $t = \hat{f}_t(z)$ (Fig. 3.5-left) and when regressing the altitude on the temperature one obtains the model $z = \hat{f}_z(t)$ (Fig. 3.6-left).

Interestingly enough, the expected mean squared error on the train and test sets averaged over 100 runs are lower for the *false* causal orientation $T \rightarrow Z$ than for the *true* causal orientation $Z \rightarrow T$. Therefore the overall predictive accuracy measured in term of mean squared error is better, even if the model $z = \hat{f}_z(t)$ does not seem to accurately reproduce the data generative process notably in the non-reversible part of the relation (circled area on Fig. 3.6-left).

The best predictive model does not necessarily corresponds to the true causal orientation [29]. It comes from the fact that minimizing a predictive score such as an expected mean squared error does not give necessarily an explanation of the data

[3] *GaussianProcessRegressor* algorithm with default parameters from python library scikit-learn 0.19.1 [23] are used.

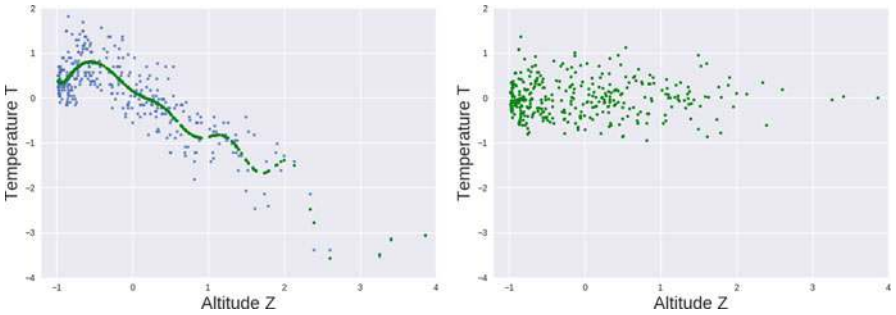

Fig. 3.5 Left: real couples of points (altitude, temperature) in blue and regression model $t = \hat{f}_t(z)$ in green. To some extent this residual is independent of the altitude Z. Therefore, the causal hypothesis $T := \hat{f}_T(Z) + N_T$ with $N_T \perp\!\!\!\perp T$ may hold

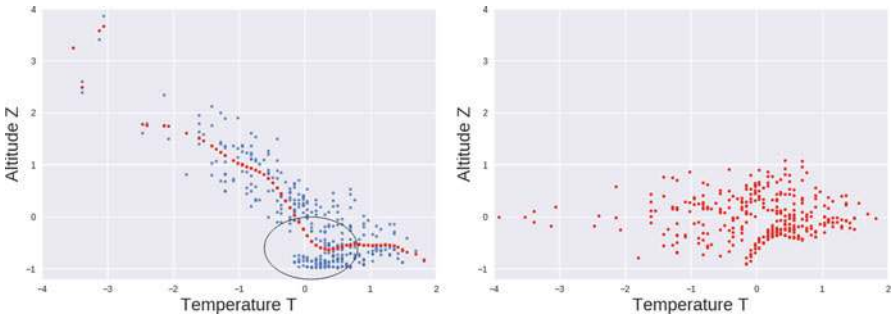

Fig. 3.6 Left: real couples of points (altitude, temperature) in blue and regression model $z = \hat{f}_z(t)$ in red. Right: residual of the non-linear regression of the altitude on temperature (red). The circled area corresponds to the non-reversible part of the relation (Z,T). To some extent this residual is not independent of the temperature T. Therefore, the causal hypothesis $Z := \hat{f}_Z(T) + N_Z$ with $N_Z \perp\!\!\!\perp T$ does not hold

generative process and then may lead to misunderstandings when it comes to causal interpretation.[4]

If one uses the same nonlinear regression model, but rather than looking at the mean squared error on the test set, one looks at the residuals on the test set defined respectively as $n_t(z) = t(z) - \hat{f}_t(z)$ and $n_z(t) = z(t) - \hat{f}_z(t)$, one can see that a causal footprint may be detected. Indeed, the residual N_Z is almost independent of Z (Fig. 3.5-right), while the residual N_T is not independent of T (Fig. 3.6-right).

[4]Let us note however that a recent work of [2] shows that comparing mean square error after fitting regression models in both direction can achieve overall good results when specific assumptions are satisfied such as the function ϕ that represents the causal mechanism is monotonically increasing (or decreasing) and a specific independence postulate between the variance of the noise and the derivative ϕ' is satisfied (see Sect. 3.5.3.4 for a description of this method). A wide comparative evaluation of all the methods, including this method RECI, will also be proposed in Sect. 3.6.

It confirms, to some extent, that the causal hypothesis $T := \hat{f}_T(Z) + N_T$ with $N_T \perp\!\!\!\perp Z$ holds, while the causal hypothesis $Z := \hat{f}_z(T) + N_Z$ with $N_Z \perp\!\!\!\perp T$ does not hold. Therefore, from an explanatory perspective, using the same conclusion as in the previous example, the additive noise model $T := \hat{f}_T(Z) + N_T$ (model 1) is preferred to the other explanation $Z := \hat{f}_Z(T, N_Z)$ (model 2) with non additive noise.

From a probabilistic point of view, the model $T := \hat{f}_T(Z) + N_T$ corresponds to the stochastic model $P_{T|Z}$ that accounts for uncertainty about the mechanism involved. Indeed for a given altitude there are several values of temperature possible, because the temperature may depend on other unobserved factors such as the latitude, the type of vegetation, the type of soil, the degree of humidity in the air, etc. In order to characterize a full *generative model*, one may consider that the altitude depends from other unknown variables such as the variations in terrain elevation: $Z := \hat{f}_Z(N_Z)$, where N_Z models latent source causes. It gives a distribution of P_Z, which when combined with $P_{T|Z}$ gives a full generative model $P_Z P_{T|Z} = P_{Z,T}$, that can be used as a simulator to draw samples (z, t) in the region.

This explanatory model may recover a causal interpretation, formally defined with the *do-notation* [22]. The temperature T is said to be a cause of Z if:

$$P_T^{\mathrm{do}(Z=z)} \neq P_T^{\mathrm{do}(Z=z')}, \tag{3.4}$$

for some z, z' [20]. An intervention, denoted as $\mathrm{do}(Z = z)$, forces the variable altitude Z to take the value z, while the rest of the system remains unchanged. Concretely this mathematical formulation can be translated into: "*all other things being equal*", when modifying the altitude (climbing a mountain), it has an impact on the temperature (it decreases)". However $P_Z^{\mathrm{do}(T=t)} = P_Z$ as modifying the temperature (heating the air) does not increase the altitude. Nevertheless this causal implication would not be true for hot air balloons! Indeed in causality "random variables" cannot be isolated from their context, because they are intimately related to an underlying specific system.

Moreover this functional causal model could also be used to derive counterfactual statements [22]. Indeed, for any specific meteorological station with a couple of datapoints altitude and temperature, (z_i, t_i), if one knows f_T, one can calculate the value n_T^i such that $t_i = f_T(z_i) + n_T^i$, and therefore for any specific station, one can answer the question "what would have been the temperature t_i' in this meteorological station if the altitude had changed from z_i to z_i'?" by using the mathematical expression $t_i' = f_T(z_i') + n_T^i$. This counterfactual reasoning on specific individuals would not be possible when having only a model $P_{T|Z}$ at the population level and not the underlying functional causal model including both causal orientation $(Z \to T)$ and mechanism f_T.

3.1.3 Comparing Alternative Data Generating Models

As shown with this real introductory example, finding the causal direction, when assuming that there are no confounding effects, consists in comparing two alternative data generating models and deciding whether the causal process $Z \to T$ is more natural or simpler than the backward process $T \to Z$. Intuitively, we can think of it as a rudimentary physical model that generates the temperature (effect) from the altitude (cause), which provides a better explanation in some way (more natural or simpler) than generating the altitude from the temperature.

In order to compare these alternatives models, an *Occam's razor principle* is always invoked in one way or another in the literature. Generally speaking, an *Occam's razor principle* can be seen as a general heuristic used in science to guide the modeler to find the *simplest explanation* when *testing different causal hypotheses on the data*. In order to apply this principle, two things must be defined : what do we mean by *simplest explanation*, which refers to the notion of the **complexity of a model**? And what do we mean by *testing a causal hypothesis on the data*, which refers to the notion of the **fit of a model**? These two notions of complexity and model fit have been formalized in different ways in the literature. We will detail them in this chapter.

Furthermore, one has always to keep in mind however that this heuristic choice is not an irrefutable principle. It is impossible in the cause-effect pair problem from purely observational data to formally prove that an explainable causal model is true. It is easy to find examples where *Occam's razor principle* favors the wrong theory given available data. Indeed in the introductory example with altitude and temperature, the *true causal mechanism* could have been $Z := f_Z(T, N_Z)$ (as shown later on in this chapter one can always exhibit such mechanism f_Z and variable N_Z with $N_Z \perp\!\!\!\perp T$) and a conclusion based on the idea that $T := f_T(Z) + N_T$ is simpler would have led to a false conclusion.

However this Occam razor principle has been implemented in the literature with good empirical success on artificial and real data [20]. By looking at the overall picture, we can distinguish three types of methods implementing this principle in different ways. The first class of methods uses fixed complexity of models and chooses the causal direction corresponding to the model that best fits the data. A second type of methods evaluates a weighted aggregation between two criteria: complexity and fit of the model. The last approaches exhibit two candidate models that are assumed to perfectly correspond to the data generative process and compare their complexities.

In Sect. 3.2, we introduce the bivariate problem setting with the usual assumptions invoked. In Sect. 3.3 we discuss the specific problem of identifiability that appears in this problem. The following Sect. 3.4 is devoted to the general method developed in the literature to tackle this identifiability issue. It will allow us to define a typology of the cause-effect inference methods that we will present more in detail in Sect. 3.5 with their practical implementations. In Sect. 3.6 we propose a benchmark of various methods presented in this chapter on artificial and real data.

The next Sect. 3.7 is a discussion on open problems and extensions for the cause-effect pair setting. The last Sect. 3.8 concludes.

3.2 Problem Setting

In this section, we present the cause-effect pair problem from the generative approach perspective. We first introduce the notations used in the chapter as well as the main assumptions usually involved. Then we present the general bivariate structural model.

3.2.1 Notations and Assumptions

X and Y are two one-dimensional random variables in $\mathbb{R}$ with joint distribution $P_{X,Y}$.

3.2.1.1 Identically and Distributed Samples

The given observations $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$ of the random variables X and Y are independent and identically distributed drawn from $P_{X,Y}$.

3.2.1.2 Time

The time for which the observed data have been collected is not available. It is then impossible to exploit Granger causality tests for time series relying on the principle that if $X \to Y$, then the predictions of the value of Y based on its own past values and on the past values of X are better than predictions of Y based only on its own past values [8]. In the approaches presented in this chapter, time is not explicitly modeled, even though it is assumed that causes precede their effects.

3.2.1.3 Faithfulness Assumption

This is the classical faithfulness assumption used in graphical causal inference, transposed for two variables: "if there is a causal relation between X and Y, the two variables are not independent".

Pathological cases could arise for example as depicted on Fig. 3.7. The altitude has negative effect on temperature, but the altitude could have also negative impact on the degree of humidity in the air, which could have a negative effect on the temperature (as strange as it may seem, but this is an illustrative case). In this

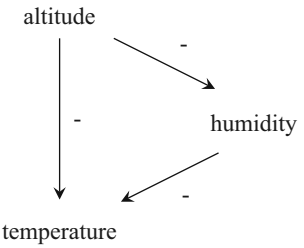

Fig. 3.7 Example of unfaithful case

scenario the altitude would directly inhibit the temperature and indirectly improve it, which would result in a statistical independence between altitude and temperature by coincidence even if it is well known that altitude causes temperature.

In this chapter we will consider only pair effect problems where the two variables X and Y are statistically dependent. When detecting a dependency between X and Y five main cases may arise:

1. X causes Y
2. Y causes X
3. Selection bias: there is a unobserved common effect W of both X and Y on which X and Y were conditioned during the data acquisition. This selection bias creates a spurious dependency between X and Y.
4. Confounder: X and Y are both common consequences of a same third variable $X \leftarrow Z \rightarrow Y$
5. Feedback loop: X is a cause of Y and Y is a cause of X.
6. Constraint relation: X and Y are linked together, but there is no causal relationship between them.

Let us note that multiple combinations such as case 1 and 4 may arise at the same time, X causes Y *and* both variables are also caused by an unobserved variable Z.

3.2.1.4 Selection Bias

A selection bias corresponds to an unobserved variable on which the two variables X and Y were implicitly conditioned: graph $X \rightarrow W \leftarrow Y$ (Fig. 3.8-(3)). In our introductory example, we may imagine for example that the variable temperature is in fact independent of the variable altitude. However the temperature has an impact on the number of dwarf mountain pines present in the region because this type of trees grows only in cold region. The altitude has also an impact of this type of vegetation as this type of trees grows only in mountains. Then if the weather stations were only constructed in area with this type of vegetation (as strange as it may seem), when collecting the data, an artificial dependency link would be created between altitude and temperature (Fig. 3.9).

Throughout this chapter, it is assumed that the sample $\{(x_i, y_i)\}_{i=1}^{n}$, corresponding to the variables X and Y, was collected without selection bias.

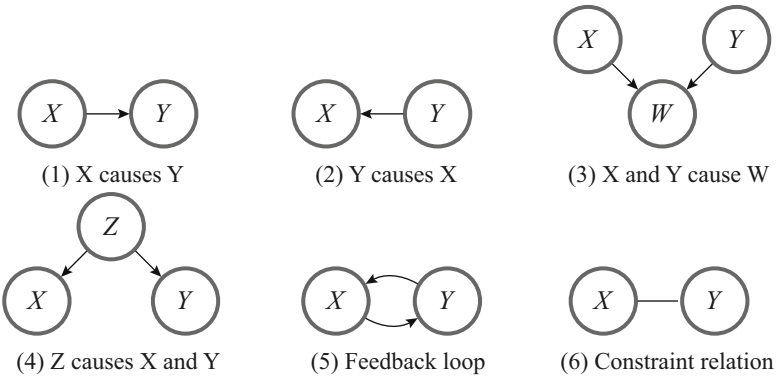

Fig. 3.8 Five main cases when detecting a dependency between X and Y

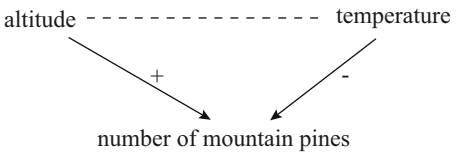

Fig. 3.9 Selection bias when measuring altitude and temperature

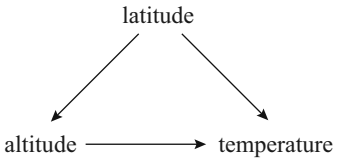

Fig. 3.10 Latitude causes altitude and temperature

3.2.1.5 Causal Sufficiency Assumption

Under this assumption it is assumed that X and Y are not common consequences of the same hidden variables (case corresponding to $X \leftarrow Z \rightarrow Y$ (Fig. 3.8-(4)) excluded).

This causal sufficiency assumption is made by many methods of the cause-effect pair literature that we will review in this chapter as it allows to considerably simplify the cause-effect pair problem. However there are many realistic cases where there are potential confounding effects that can affect both variables (the most typical examples of confounding factors are the age or the gender in epidemiological studies). In this regard, if we come back to our introductory example, a confounding effect may be present as illustrated on Fig. 3.10. Indeed, the hidden variable latitude is known to be a cause of the temperature, but it is also a cause of the altitude, because in Germany all the mountains are situated in the south of the country. Therefore the link between altitude and temperature could be completely spurious and disappears when conditioned on the latitude variable. However it is not the case with this example as we still have a well known causal effect from altitude to temperature, but it highlights the fact that for the cause-effect pair analysis, one

should always study the relationship between X and Y after conditioning on the potential observed confounding variables. In Sect. 3.7, we will provide a discussion on the confounding case, which is a great challenge and the problem is far from being solved for the cause-effect pair setting. To keep the problem simple in the first instance and clearly explain the cause-effect pair problem where only $X \to Y$ or $Y \to X$ are considered, we will first assume in this chapter that the confounding case is excluded.

3.2.1.6 Feedback Loops

A feedback loop appears when both X causes Y and Y cause X: graph $X \rightleftharpoons Y$ (Fig. 3.8-(5)). Even if the notion of time is not present, this case may happen for example in cross-sectional studies where data are collected over a certain period of time. Mooij et al. [20] give an example where the two variables are the temperature and the amount of sea ice: "an increase of the global temperature causes sea ice to melt, which causes the temperature to rise further (because ice reflects more sun light)".

We will assume in this chapter that this case is excluded. Therefore the causal graph between X and Y refers to the literature on Directed Acyclic Graph (DAG) (with only two variables).

3.2.1.7 Constraint Relation

This case of dependency with a relation $X-Y$ (Fig. 3.8-(4)) may arise for example if the two variables are linked by a logical or mathematical formula, and not related by a causal relation. As an example, if we write that the productivity P in a firm is equal to the total added value VA of the firm divided by the total number of employees N, the two variables P and VA are linked by a mathematical expression: $P = \frac{VA}{N}$. By observing VA and knowing N we can immediately deduce P, but the opposite is also true as knowing P and N gives VA, with equivalent mathematical relation, without any notion of an underlying system that could have generated one of the variable from the other. We do not consider this case in this chapter.

3.2.1.8 Measurement Noise

In general in the literature, it is assumed that there are no measurement noise. Such measurement noise may happen if for example the altitude is not measured precisely, but its noisy version noted $\tilde{Z}$. However the variable temperature T is still a function of the original variable Z that is not corrupted by measurement noise. This problem is related to a cause-effect pair problem between T and $\tilde{Z}$ in presence of a latent hidden variable which is the original Z. We refer the reader to [20] who propose a

benchmark that verify the robustness of various cause-effect algorithms in presence of small perturbations of the data.

3.2.1.9 Variables Units

The orders of magnitude of the variable as well as their physical units are not considered in the problem. In most methods, the variables X and Y are re-scaled with zero mean and unit variance.

3.2.2 Bivariate Functional Causal Models

In the literature on causality, if $\mathcal{G}$ denotes an acyclic causal graph (DAG) obtained by drawing arrows from causes $X_{\mathrm{Pa}(i;\mathcal{G})}$ towards their effects X_i, it is often assumed that the effects are expressed as a linear function of their cause and an additive Gaussian noise. These models are linear structural equation models, where each variable is continuous and modeled as:

$$X_i := \sum_{j \in \mathrm{Pa}(i;\mathcal{G})} \alpha_j X_j + N_i, \text{ for } i = 1, \ldots, d, \tag{3.5}$$

with $\mathrm{Pa}(i;\mathcal{G})$ the subset of index of the parents of each variable X_i in graph $\mathcal{G}$ and N_i a random noise variable, meant to account for all unobserved variables. The parameters α_j are real values. Each equation characterizes the direct causal relation explaining variable X_i from the set of its causes $X_{\mathrm{Pa}(i;\mathcal{G})}$, based on some linear *causal mechanisms*. These models are used a lot in social science fields such as econometric and sociology. Although this simplified model with linear mechanisms and additive Gaussian noise appears to be very convenient from a theoretical point of view, it is not often realistic as the interactions between cause and noise may be more complex in reality. Therefore a more general framework has been proposed by Pearl [21] with potential nonlinear interactions between cause and effect:

$$X_i := f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, N_i), \text{ for } i = 1, \ldots, d. \tag{3.6}$$

When the DAG $\mathcal{G}$ is reduced to $X \to Y$, this system of equation refers to a bivariate structural model.

Definition 3.1 A bivariate structural model noted $\mathcal{B}_{\mathcal{G}, f, P_N}$ is a triplet $(\mathcal{G}, f, P_N)$, where $\mathcal{G}$ is the causal graph $X \to Y$ or $Y \to X$, $f = (f_X, f_Y)$ is a couple of possibly nonlinear functions and (N_X, N_Y) are two independent random variables drawn according to continuous distribution $P_N = (P_{N_X}, P_{N_Y})$, such that:

- $X := f_X(N_X)$ and $Y := f_Y(X, N_Y)$ if $X \to Y$
- $Y := f_Y(N_Y)$ and $X := f_X(Y, N_X)$ if $Y \to X$

One can notice that this definition holds for any type of continuous distribution of the noise P_N. For example P_{N_Y} can be set to the uniform distribution on [0, 1] and this is not a general restriction, since one can always write $N_Y = g_Y(\tilde{N}_Y)$, for some function g_Y, with $\tilde{N}_Y \sim \mathscr{U}[0, 1]$ and $\tilde{f}_Y = f_Y(\cdot, g_Y(\cdot))$ [31].

Without loss of generality, we could also write $X := N_X$, with $P_{N_X} = P_X$, but in the following we prefer to keep the formulation with two functions (f_X, f_Y) in order to stay consistent with the general formulation of FCM given by Eq. (3.6).

According to [20], the assumption that N_X and N_Y are independent, is justified by the assumptions that there is no confounding effect, no selection bias, and no feedback between X and Y (see Sect. 3.2.1).

3.3 The Problem of Identifiability with Two Variables

Given the formulation of the processes described in Sect. 3.2.2, the task is to identify the causal structure $X \to Y$ or $Y \to X$ that could have generated the observed data. By *identifying* we mean proving that f and P_N exist so that the hypothesis $\mathscr{B}_{\mathscr{G},f,P_N}$ holds in the causal direction $\mathscr{G}$ with respect to the observed data while there do not exist any f' and P'_N so that $\mathscr{B}_{\mathscr{G}',f',P'_N}$ holds in the opposite causal direction $\mathscr{G}'$.

This problem faces two difficulties. The first is a classical empirical problem because in general one has access to a finite sample size $\mathscr{D} = \{(x_j, y_j]\}_{j=1}^n$ making impossible to evaluate perfectly $P_{X,Y}$. This is why the evaluation will rely on the definition of a model $Q_{X,Y}$ of the data distribution, which we will discuss later on. The second difficulty is more profound as it is related to the inference of a DAG when only two variables are observed. In this case, it is impossible to identify the causal direction by using classical conditional independence tests (e.g. as in the PC algorithm of [30]), because the two graphs $X \to Y$ and $Y \to X$ are Markov equivalent.

3.3.1 The Particular Linear Gaussian Case

A first well known identifiability issue arises in the linear Gaussian case because it induces a perfectly symmetric distribution after rescaling (Fig. 3.11).

A linear Gaussian generative bivariate FCM is defined by the system of equations:

$$\begin{cases} X := \alpha_X N_X \text{ with } N_X \sim \mathscr{N}(\mu_X, \sigma_X) \\ Y := \beta_Y X + \alpha_Y N_Y, \text{ with } N_Y \sim \mathscr{N}(\mu_Y, \sigma_Y), \end{cases} \tag{3.7}$$

with $\alpha_X, \alpha_Y, \beta_Y \in \mathscr{R}^3$. As shown by Mooij et al. [20], it is always possible in this case to find parameters $\alpha'_X, \alpha'_Y, \beta'_X, \mu'_X, \mu'_Y, \sigma'_X, \sigma'_Y$ such that:

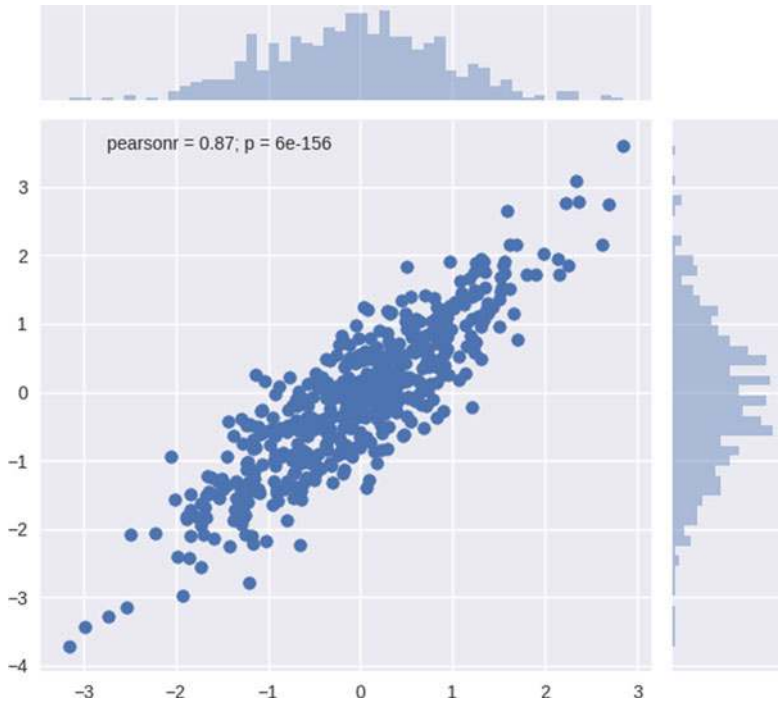

Fig. 3.11 Pairwise linear Gaussian case

$$\begin{cases} Y := \alpha'_Y N_Y \text{ with } N_Y \sim \mathcal{N}(\mu'_Y, \sigma'_Y) \\ X := \beta'_X Y + \alpha'_X N_X, \text{ with } N_X \sim \mathcal{N}(\mu'_X, \sigma'_X). \end{cases} \tag{3.8}$$

Therefore, for this pair a perfect symmetric generative model exists in both directions (that can only be dissociated by the values of its parameters) and it is impossible to determine the causal direction from these observational data.

3.3.2 General Case

Given any two random variables X and Y with continuous support, [36] shows that if $F_{Y|X}$ is the conditional cumulative distribution function of Y given X and q an arbitrary continuous and strictly monotonic function with a non-zero derivative, then the quantity $\tilde{N} = q \circ F_{Y|X}$, where $\circ$ denotes function composition is independent from X. Furthermore, the transformation from (X, Y) to $(X, \tilde{N})$ is always invertible, in the sense that Y can be uniquely reconstructed from $(X, \tilde{N})$.

Stated in another way, given any two random variables X and Y with continuous support, one can always construct a function f_Y and another variable, denoted by

N_Y, which is statistically independent from X and such that $Y := f_Y(X, N_Y)$. And equivalently, on can always construct a function f_X and another variable, denoted by N_X, which is statically independent from Y and such that $X := f_X(Y, N_X)$.

3.3.2.1 Characterization with Continuous Quantile Functions and Uniform Noise

More specifically if the joint density function h of $P_{X,Y}$ is continuous and strictly positive on a compact and convex subset of $\mathbb{R}^2$, and zero elsewhere, the model $\mathscr{B}_{\mathscr{G},f,P_N}$ holds with a couple of continuous quantile functions $f = (f_X, f_Y)$ in any direction $X \to Y$ and $Y \to X$.

Indeed, if one considers the cumulative distribution F_X defined over the domain of X ($F_X(x) = Pr(X < x)$). F_X is strictly monotonous as the joint density function is strictly positive therefore its inverse, the quantile function $Q_X : [0, 1] \mapsto dom(X)$ is defined. If n_X is drawn in $\mathscr{U}[0, 1]$, by construction, $Q_X(n_X) = F_X^{-1}(n_X)$ and by setting $f_X = Q_X$, we obtain $X = f_X(N_X)$. For any noise value n_X let x be the value of X defined from n_X. The conditional cumulative distribution $F_Y(y|X = x) = Pr(Y < y|X = x)$ is strictly continuous and monotonous with respect to y, and can be inverted using the same argument as above. Then we can define $f_Y(x, n_y) = F_Y^{-1}(x, n_y)$ and we obtain $Y := f_Y(X, N_Y)$. In an equivalent manner, we can show that there exists a set $f = (f_X, f_Y)$ such that $Y := f_Y(N_Y)$ and $X := f_X(Y, N_X)$.

Furthermore, it has been shown by Goudet et al. [7] that under the same assumptions on $P_{X,Y}$ for both candidate generative bivariate FCM $X \to Y$ and $Y \to X$, the functions f_X and f_Y defined above are continuous on their definition domain.

An example of a continuous joint density function $P_{X,Y}$, strictly positive on a compact and convex subset of $\mathbb{R}^2$ and zero elsewhere is depicted on Fig. 3.12. On

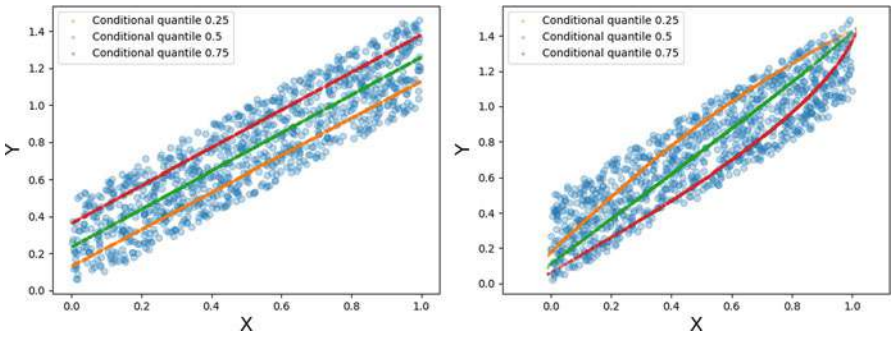

Fig. 3.12 Cause-effect pair with quantile regressions 0.25 (orange), 0.5 (green) and 0.75 (red) in both directions $X \to Y$ (left) and $Y \to X$ (right). Better seen in color

the left part of the Figure is shown a quantile regression of the variable Y on X with fraction 0.25 (orange curve), 0.5 (green curve) and 0.75 (red curve). It corresponds to the estimation of the FCM $y := f_Y(x, n_y)$ with uniform noise and where n_y is respectively set to the values 0.25, 0.5 and 0.75. On the right part of this Figure is shown the quantile regression of X on Y, corresponding to the estimation of the FCM $x := f_X(y, n_x)$, with n_x set to the values 0.25 (orange curve), 0.5 (green curve) and 0.75 (red curve). It highlights the fact that continuous FCM may be recovered in both directions in this case. However for the causal orientation $X \to Y$, the FCM is linear for any fixed noise value n_y, while for the causal orientation $Y \to X$, the FCM is more complicated as it seems to be only linear for $n_x = 0.5$ (green curve).

3.3.2.2 How to Overcome This Identifiability Problem?

This identifiability problem is a negative result for the cause-effect pair problem, because without any additional assumptions the problem is unsolvable.

However, even if both $\mathscr{B}_{X \to Y, f, P_N}$ and $\mathscr{B}_{Y \to X, f', P'_N}$ hold, there is almost always an asymmetry in the data generative process $X \to Y$ and $Y \to X$, because in general the mechanisms f and f' do not belong to the same class of functions (except in the linear Gaussian case mentioned before).

If we go back to the introductory example with altitude and temperature, there exist two plausible causal models:

$$\begin{cases} Z := f_Z(N_Z) \\ T := f_T(Z, N_T), \ with \ N_T \perp\!\!\!\perp Z \end{cases} \tag{3.9}$$

$$\begin{cases} T := f_T(N_T) \\ Z := f_Z(T, N_Z), \ with \ N_Z \perp\!\!\!\perp T \end{cases} \tag{3.10}$$

However, the first model of Eq. (3.9) can be rewritten *to some extent* with an additive mechanism:

$$\begin{cases} Z := f_Z(N_Z) \\ T := f_T(Z) + N_T, \ with \ N_T \perp\!\!\!\perp Z, \end{cases} \tag{3.11}$$

while the alternative causal model with causal orientation $T \to Z$ cannot be expressed using the same type of expression with additive noise. If one accepts the fact that an additive mechanism is a "simpler" form of conditional, we may prefer the causal orientation $Z \to T$ according to Occam's Razor principle.

In general, one can see that the factorization of the joint density function $P_{X,Y}$ into $P_X P_{Y|X}$ or $P_Y P_{X|Y}$ may lead to models with different complexities, with respect to some appropriate notion of complexity to be defined.

3.4 Computing a Trade-Off Fit/Complexity

The determination of the best explainable model is then based in the literature on these main lines:

- Different candidate bivariate models (hypotheses) are evaluated in both directions.
- For each candidate model one evaluates a score monitoring the trade-off between the fit of the model (meaning its adequacy to the observational data) and the complexity of the mechanisms involved.
- The model with the best score is returned, with its corresponding causal arrow $X \to Y$ or $Y \to X$.

3.4.1 Defining Candidate Bivariate Models and Sampling Data

In order to model such continuous underlying bivariate generative process $\mathscr{B}_{\mathscr{G},f,P_N}$ which is assumed to have generated the data, we introduce the notion of candidate model $\widehat{\mathscr{B}}_{\widehat{\mathscr{G}},\hat{f},Q_N}$ described by:

- a structure defined by a causal orientation $X \to Y$ or $Y \to X$
- its estimated mechanisms modeled by $\hat{f}$
- an estimated distribution of noise Q_N

3.4.1.1 Structure of a Candidate Model

We note θ_X and θ_X the vectors of parameters of the estimated mechanisms $\hat{f}_X$ and $\hat{f}_Y$. And we note respectively θ_{N_X} and θ_{N_Y} the vectors of parameters of the distribution of the modeled noise variables Q_{N_X} and Q_{N_Y}. The noise variables are independent. The global vector of parameters of the model is noted $\theta = (\theta_X, \theta_Y, \theta_{N_X}, \theta_{N_Y})$.

When $\widehat{\mathscr{G}} = X \to Y$, this candidate generative model depicted on Fig. 3.13 generates a distribution $Q_{X,Y}(\theta) = Q_X(\theta_X, \theta_{N_X})Q_{Y|X}(\theta_Y, \theta_{N_Y})$. When $\widehat{\mathscr{G}} = Y \to X$, this candidate model generates a distribution $Q_{X,Y}(\theta) = Q_Y(\theta_Y, \theta_{N_Y})Q_{X|Y}(\theta_X, \theta_{N_X})$.

In some approaches proposed in the literature, the cause variable is not modelled but taken as the observed variable. In this case, $Q_X = P_X$. We refer the reader to Chap. 1 of this book explaining why it could be a problem for cause-effect inference.

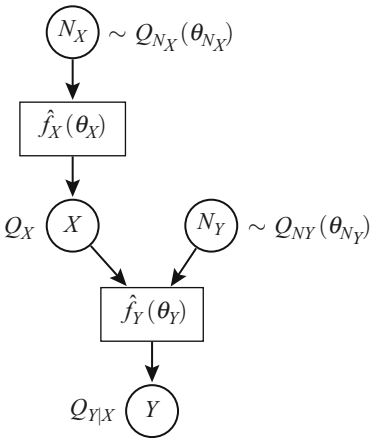

Fig. 3.13 Candidate generative model with causal orientation $X \to Y$

3.4.1.2 Mechanisms

After having defined the overall structure, one needs to model the mechanisms:

- A first characterization that needs to be specified is related to the type of interaction between the noise variable and the cause. Indeed in the literature, we distinguish mainly *additive noise interaction* of the form $Y := \hat{f}_Y(X) + N_Y$ or *complex noise interaction* of the form $Y := \hat{f}_Y(X, N_Y)$, where the cause variable and the noise are mixed with a non-linear mechanism.
- The second characterization concerns the class of functions used to define $\hat{f}$. It may range from linear mechanisms as in LiNGAM algorithm [28] to complex non-linear mechanisms modeled with Gaussian processes [31] or neural networks [16]. In general, the more complex the mechanisms are, the more the candidate model can fit the data and the more the model is general (meaning that it can be applied to a wide variety of cases). However, it may result in more difficulty to assess the causal orientation because the candidate model has more chances to fit equally well the data in both directions. A method for controlling the complexity of the mechanisms involved will be discussed later on in Sect. 3.4.3.

3.4.1.3 Sampling Data Points with a Candidate Model

Now we have all the ingredients required to sample data points of the estimated distribution $Q_{X,Y}$ with the **generative model** depicted in Fig. 3.13 by proceeding as follow:

1. Draw $\{(n_{X,j}, n_{Y,j})\}_{j=1}^{n}$, n samples independent and identically distributed from the joint distribution $Q_{N_X}(\theta_{N_X}) \times Q_{N_Y}(\theta_{N_Y})$ of independent noise variables (N_X, N_Y).

2. Generate n samples $\widehat{\mathscr{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{n}$, where each estimated sample $\hat{x}_j$ of variable X is computed from $\hat{f}_X(\theta_X)$ with the j-th estimated samples $n_{X,j}$; then each estimated sample $\hat{y}_j$ of variable Y is computed from $\hat{f}_Y(\theta_Y)$ with the j-th estimated samples $n_{Y,j}$ and $\hat{x}_j$.

Generative candidate models support interventions, that is, freezing the variable X to some constant v_i. The resulting joint distribution noted $Q_Y^{\mathrm{do}(X=v_i)}$, called *interventional distribution* [22], can be computed from this model by clamping the value of X to the value v_i. However when freezing the value of Y to some constant w_i, it has no impact on X: $Q_X^{\mathrm{do}(Y=w_i)} = Q_X$. Generative candidate models support also counterfactual reasoning [22] as explained with the introductory example.

3.4.2 Model Fitness Score

In order to evaluate the quality of a candidate generative model, we introduce a fit score $S_{\widehat{\mathscr{B}}}(\theta)$ of a candidate model $\widehat{\mathscr{B}}_{\widehat{\mathscr{G}},\hat{f},Q_N}(\theta)$. This score has been implemented in different ways in the literature. It has always the property to be minimal when $P = Q$ (perfect fit in the large sample limit).

3.4.2.1 Log-Likelihood Parametric Scores

An example of score used in [31, 32] is the negative log-likelihood score defined for a candidate generative model with causal orientation $X \to Y$ by:

$$\begin{aligned} S_{\widehat{\mathscr{B}}}(\theta) &= -\log \hat{f}(\mathscr{D}|\theta) \\ &= -\sum_{i=1}^{n} \left[\log Q_{X=x^i}(\theta_X, \theta_{N_X}) + \log Q_{Y=y^i|X=x^i}(\theta_Y, \theta_{N_Y})\right]. \end{aligned} \tag{3.12}$$

This likelihood score is often computed in a parametric context with special constraints imposed on the class of densities for the distribution of the cause Q_X and the distribution of the conditional $Q_{Y|X}$. For example in [31], a Gaussian mixture model is used as a prior distribution of the cause and a Gaussian process with a zero mean function and a squared-exponential covariance function is chosen as prior of the conditional.

3.4.2.2 Implicit Fit Score Computed as an Independence Score Between Cause and Noise Variable

If one considers a model with causal orientation $X \to Y$, [36] shows that computing this maximum likelihood score of the model with respect to the observational data is equivalent to minimizing a mutual information term $I(X, \hat{N}_Y; \theta)$ between the estimated noise $\hat{N}_Y$ and the cause X. We re-transcribed here the theoretical justification given by Zhang et al. [36] with our notations.

We consider a candidate model $\widehat{\mathscr{B}}_{\widehat{\mathscr{G}}, \hat{f}, Q_N}(\theta)$ with causal orientation $X \to Y$ and where the distribution of the source cause is not modelled and taken as P_X.

One can write $Q_{X,N_Y} = P_X Q_{N_Y}$ as in this model X and N_Y are assumed to be independent. Therefore, the Jacobian matrix of the transformation from (X, N_Y) to (X, Y) is:

$$\mathbf{J}_{X \to Y} = \begin{pmatrix} \frac{\delta X}{\delta X} & \frac{\delta X}{\delta N_Y} \\ \frac{\delta Y}{\delta X} & \frac{\delta Y}{\delta N_Y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{\delta f_Y}{\delta X} & \frac{\delta f_Y}{\delta N_Y} \end{pmatrix}. \tag{3.13}$$

The absolute value of its determinant is $|\mathbf{J}_{X \to Y}| = |\frac{\delta f_Y}{\delta N_Y}|$. Then,

$$Q_{X,Y} = Q_{X,N_Y} / |\mathbf{J}_{X \to Y}| = P_X Q_{N_Y} |\frac{\delta f_Y}{\delta N_Y}|^{-1}, \tag{3.14}$$

which implies

$$Q_{Y|X} = Q_{N_Y} |\frac{\delta f_Y}{\delta N_Y}|^{-1}. \tag{3.15}$$

As the transformation from (X, N_Y) to (X, Y) is invertible, given any parameter set θ_Y involved in the function f_Y, the noise N_Y can be recovered, and the authors denote by $\hat{N}_Y$ the estimated noise variable. For any parameter set $\theta = (\theta_Y, \theta_{N_Y})$, using Eq. (3.15) the negative log-likelihood score attained by the model is:

$$\begin{aligned} S_{\widehat{\mathscr{B}}}(\theta) &= -\log \hat{f}(\mathscr{D}|\theta) \\ &= -\sum_{i=1}^{n} \left[\log P_{X=x^i} + \log Q_{Y=y^i|X=x^i}(\theta_Y, \theta_{N_Y})\right] \\ &= -\sum_{i=1}^{n} \log P_{X=x^i} - \sum_{i=1}^{n} \log Q_{N_Y=\hat{n}_Y^i}(\theta_{N_Y}) \\ &\quad + \sum_{i=1}^{n} \log |\frac{\delta f_Y}{\delta N_Y}_{X=x^i, N_Y=\hat{n}_Y^i}(\theta_Y)|. \end{aligned} \tag{3.16}$$

Now the fit of the model can be seen differently. Instead of fitting the model (Sect. 3.2.2) by modeling the noise N_Y which is independent of X and then modeling the conditional $Q_{Y|X}$, one can start from the true distribution $P_{X,Y}$ and look for such an estimate $\hat{N}_Y$ that $\hat{N}_Y$ and X are independent. In this approach, (X, Y) is recovered from $(X, \hat{N}_Y)$ as:

$$P_{X,Y} = Q_{X,\hat{N}_Y} |\frac{\delta f}{\delta N_Y}|^{-1}. \tag{3.17}$$

In order to make $\hat{N}_Y$ and X independent, one can compute the mutual information between X and $\hat{N}_Y$ that depends of the parameters θ of the model:

$$\begin{aligned} I(X, \hat{N}_Y; \theta) &= \mathbb{E}_{x\sim P_X, \hat{n}_Y \sim Q_{\hat{N}_Y}} \left[\log \frac{Q_{X=x,\hat{N}_Y=\hat{n}_Y}}{P_{X=x} Q_{\hat{N}_Y=\hat{n}_Y}} \right] \\ &= -\mathbb{E}_{x\sim P_X} \log P_{X=x} - \mathbb{E}_{\hat{n}_Y \sim Q_{\hat{N}_Y}} \log Q_{\hat{N}_Y=\hat{n}_Y} \\ &\quad + \mathbb{E}_{x\sim P_X, \hat{n}_Y \sim Q_{\hat{N}_Y}} \log Q_{X=x,\hat{N}_Y=\hat{n}_Y}. \end{aligned} \tag{3.18}$$

By using the sample version of this quantity and Eq. (3.17) the authors obtain:

$$\begin{aligned} \hat{I}(X, \hat{N}_Y; \theta) = &-\frac{1}{n}\sum_{i=1}^{n} \log P_{X=x^i} - \frac{1}{n}\sum_{i=1}^{n} \log Q_{\hat{N}_Y=\hat{n}_Y^i}(\theta_{N_Y}) \\ &+ \frac{1}{n}\sum_{i=1}^{n} \log |\frac{\delta f_Y}{\delta N_Y}_{X=x^i, N_Y=\hat{n}_Y^i}(\theta_Y)| + \frac{1}{n}\sum_{i=1}^{n} \log P(X=x^i, Y=y^i). \end{aligned} \tag{3.19}$$

Thus,

$$S_{\hat{\mathscr{B}}}(\theta) = n\hat{I}(X, \hat{N}_Y; \theta) - \sum_{i=1}^{n} log P(X=x^i, Y=y^i). \tag{3.20}$$

As the term $\sum_{i=1}^{n} log P(X=x^i, Y=y^i)$ does not depend on the parameters θ of the model, minimizing the likelihood score $S_{\hat{\mathscr{B}}}(\theta)$ amounts to minimizing the mutual information between the cause X and the estimated noise $\hat{N}_Y$.

This kind of evaluation score with an independence test is used to fit the model ANM [12] and PNL [35] presented in Sect. 3.5.1.

3.4.2.3 Non-parametric Scores

Other methods such as [7, 16] use a non-parametric approach. In these methods the authors do not specify any type of distribution for the model of the data $Q_{X,Y}$, but

instead compare directly with a two sample test the observed data sample $\mathscr{D}$ with the generated data sample $\widehat{\mathscr{D}}$ coming from any candidate model.

In [16], the score of the model is non-parametric and computed with Conditional Generative Adversarial Networks, or CGANs [18] (see Sect. 3.5.2.2). It has been shown by Goodfellow et al. [6] that in the large sample limit, the Generative Adversarial Networks allows to approximate the Jensen-Shannon divergence between the true distribution P and the generated distribution Q:

$$S_{\widehat{\mathscr{B}}}(\theta) \simeq JSD(P, Q). \tag{3.21}$$

The Jensen–Shannon divergence is an information theory method of measuring the similarity between two probability distributions. This metric is always positive and equal to zero when $P = Q$.

Goudet et al. [7] used the empirical Maximum Mean Discrepancy (MMD) [10], a kernel based metric between distribution, as a two sample test to compare $\mathscr{D}$ and $\widehat{\mathscr{D}}$ (see Sect. 3.5.2.3).

3.4.3 Complexity of the Model

Now we introduce a term of complexity of a candidate bivariate model $\widehat{\mathscr{B}}_{\widehat{\mathscr{G}},\hat{f},Q_N}(\theta)$ that we note $C_{\widehat{\mathscr{B}}}(\theta)$. This complexity notion refers to a simplicity prior on the underlying data generative process. It has been handled in different ways in the literature from fix a class of admissible mechanisms to more flexible criteria.

3.4.3.1 Explicit Class of Admissible Mechanisms

A first type of methods in the literature defines a rudimentary notion of complexity: all mechanisms $\hat{f}$ belonging to a particular class $\mathscr{F}$ of bivariate FCM are assumed to be simple, while the others are assumed to be complex. In [12], for a causal model with orientation $X \to Y$, a mechanism is assumed to be simple if it can be written under the form $Y := \hat{f}_Y(X) + N_Y$ with $X \perp\!\!\!\perp N_Y$ (additive noise model). A more general class of mechanism is defined by the Post-Nonlinear model (PNL) [35], involving an additional nonlinear function on the top of an additive noise: $Y := \hat{g_Y}(\hat{f}_Y(X) + N_Y)$, with $\hat{g_Y}$ an invertible function and $X \perp\!\!\!\perp N_Y$. For these methods, we write $C_{\widehat{\mathscr{B}}}(\theta) = 0$ if $\hat{f} \in \mathscr{F}$ and $C_{\widehat{\mathscr{B}}}(\theta) = 1$ otherwise.

In [7] (CGNN), the class of functional causal model is defined as $Y := \hat{f}_Y(X, N_Y)$ where $\hat{f}_Y$ is a one hidden unit neural network with RelU activation functions and $N_Y \sim \mathscr{N}(0, 1)$ with $X \perp\!\!\!\perp N_Y$. The class of admissible mechanisms is variable as it is measured as a number of hidden units n_h in this one hidden layer neural network. In this framework, the complexity term can be expressed as $C_{\widehat{\mathscr{B}}}(\theta) = n_h$.

3.4.3.2 Kolmogorov Complexity

In the previous section, the complexity term was defined in term of an explicit class of functionals, but another approach coming from the information theory has been developed in the literature.

In this information theory framework, the basic postulate that "the factorization of the joint density function $P_{\text{cause,effect}}$ into $P_{\text{cause}}P_{\text{effect|cause}}$ should lead to a simpler model than $P_{\text{effect}}P_{\text{cause|effect}}$", can be expressed with the Kolmogorov complexity framework as shown by [14]:

$$K(P_{\text{cause}}) + K(P_{\text{effect|cause}}) \overset{+}{\leq} K(P_{\text{effect}}) + K(P_{\text{cause|effect}}). \tag{3.22}$$

This inequality comes from the postulate of algorithmic independence between the distribution of the cause P_{cause} and the distribution of the causal mechanism $P_{\text{effect|cause}}$ stated by Janzing and Schölkopf [14] as:

$$I(P_{\text{cause}} : P_{\text{effect|cause}}) \overset{+}{=} 0. \tag{3.23}$$

where $I(P_{\text{cause}} : P_{\text{effect|cause}})$ denotes algorithmic mutual information.

Kolmogorov complexity and algorithmic mutual information are not computable in practice but they have led to two different practical implementations in the literature.

Model Selection with Minimum Message Length Principle (MML)

A first practical implementation of the postulate defined in Eq. (3.22) is the Minimum Message Length principle (MML), which can be seen as an information theory restatement of Occam's Razor principle.

For a candidate bivariate generative model $\mathscr{B}_{\mathscr{G},\hat{f},Q_N}$, we have defined a family of functions $\hat{f}$ (i.e. linear mechanisms, neural networks, Gaussian processes) and a family of noise distributions Q_N. The overall model (mechanisms and noise distributions) is parametrized by the vector of parameter $\theta \in \Theta$. Furthermore one can also define a prior probability distribution π of θ over Θ, that can be seen as a simplicity prior over the parameter space. Now according to the MML principle, the modeling problem of transmitting the observed data $\mathscr{D} = \{(x^i, y^i)\}_{i=1}^n$ to a receiver can be decomposed into two parts:

1. First, the model $\mathscr{B}_{\mathscr{G},\hat{f},Q_N}$ with parameter θ from the parameter space Θ is send by the transmitter to the receiver (assertion). According to [33] the complexity of this model seen in term of total code length can be approximated as:

$$C_{\mathscr{B}}(\theta) \approx -\log\pi(\theta) + \frac{1}{2}\log|\mathbf{J}_\theta| - \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi), \tag{3.24}$$

where $|\mathbf{J}_\theta|$ is the determinant of the Fisher information matrix defined as the matrix of expected second derivatives of the negative log-likelihood function with respect to all continuous parameters of the model, with entry (i, j) given by:

$$I(\theta)_{i,j} = -\int_{(x,y)\in\mathbb{R}^2} \hat{f}((x, y)|\theta)\frac{\delta^2}{\delta\theta_{i,j}}\log\hat{f}((x, y)|\theta)dxdy. \tag{3.25}$$

2. Second, the data $\mathscr{D}$ are transmitted to the receiver using this model. The length of the detail, which encodes the data using the model defined by the set of parameter θ, corresponds to the negative log-likelihood of the data according to the model defined previously in Eq. (3.12) as $S_{\hat{\mathscr{B}}}(\theta) = -\log\hat{f}(\mathscr{D}|\theta)$.

This MML principle is used for example by Stegle et al. [31] to select the model associated to $X \to Y$ or $Y \to X$ with the lower total code length $\mathscr{A}_{\hat{\mathscr{B}}}(\theta) = C_{\hat{\mathscr{B}}}(\theta) - \log\hat{f}(\mathscr{D}|\theta)$ (see Sect. 3.5.2).

Independence Between Cause and Mechanism

Another characterization of complexity comes from the algorithmic independence principle of Eq. (3.23) to derive the idea that if $X \to Y$, the marginal probability distribution of the causal mechanism $P_{Y|X}$ should be independent of the cause P_X, hence estimating a model $Q_{Y|X}$ from P_X should hardly be possible, while estimating a model $Q_{Y|X}$ based on P_X may be possible.

This complexity measure has been evaluated directly as a covariance estimation between $Q_{Y|X}$ and P_X in [4] or by a characterization of the variance of the kernel embedding of $Q_{Y|X}$ when X varies according in its definition domain [19] (see Sect. 3.5.3).

One can see that there is a direct connection between "the postulate of independence between the cause and the mechanism" and the complexity of the mechanism, as when $P_{Y|X}$ is independent on P_X, the function $\hat{f}_Y$ required to model $Q_{Y|X}$ takes in general a simpler form than the function $\hat{f}_X$ required to model $Q_{X|Y}$.

3.4.4 Bi-objective Trade-Off for Cause-Effect Inference

In the previous section, we have defined the complexity $C_{\hat{\mathscr{B}}}(\theta)$ and the fit $S_{\hat{\mathscr{B}}}(\theta)$ of a candidate bivariate model parameterized by the vector of parameters θ. From a general point of view, we can now frame the model selection as a bi-objective trade-off with a Pareto front of optimal models (cf. Fig. 3.14).

This general bi-objective trade-off between fit and complexity is very general in science. It has been seen from two different angles. Some of the modeling approaches favor models of lower complexity, while others favor models with better explanatory power (Fig. 3.14):

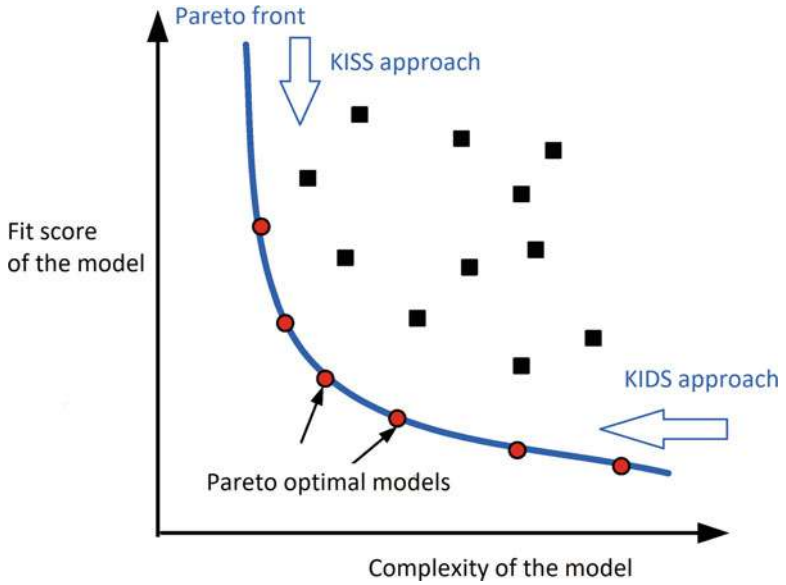

Fig. 3.14 Bi-objective trade-off between complexity and reproduction

- The first is the KISS approach (*Keep It Simple, Stupid*) proposed by Axelrod and Hamilton [1] and stating that the models should include the minimal number of parameters and mechanisms and only add new ones only if required.
- On the other hand, the KIDS principle (*Keep It Descriptive, Stupid*) is an anti-simplistic approach that favors the accuracy of the model and states that all parameters and mechanisms that appear relevant should be included, but parameters that do not add quality of the model should be removed [5].

When transposing these modeling approaches to the cause-effect problem, we can see that some methods favor models with low complexity and low explanatory power such as linear Gaussian models while others have a better explanatory power as they can model complex interactions between noise and causes (for example when using generative neural networks as in [16]), but at the cost of more complex mechanisms.

In the cause-effect pair literature, we can distinguish three different approaches used to deal with this complexity/fit trade-off:

- Methods such as those by [7, 12, 16, 35] that reason at fixed complexity $C_{\widehat{\mathscr{B}}}(\theta)$ and choose the causal direction according to the best fit. For any candidate model $\mathscr{B}_{\widehat{\mathscr{G}},\hat{f},Q_N}$, we note $\hat{\theta}$ the set of parameters that minimizes the score $S_{\widehat{\mathscr{B}}}(\theta)$ and for the purposes of simplified notations, we note $S_{X\to Y}(\hat{\theta})$ the best fit score corresponding to the model $\mathscr{B}_{\widehat{\mathscr{G}},\hat{f},Q_N}$ with $\widehat{\mathscr{G}} = X \to Y$.

If $S_{X\to Y}(\hat{\theta}) < S_{Y\to X}(\hat{\theta})$, it is decided that $X \to Y$, otherwise it is decided that $Y \to X$.

- Methods such as those by [4, 27] that assume a comparative fit $P = Q$ in both directions and choose the causal direction by comparing the complexity terms. In this case, for any candidate model $\mathscr{B}_{\widehat{\mathscr{G}},\hat{f},Q_N}$, we note $\hat{\theta}$ the set of parameters that minimizes the score $C_{\widehat{\mathscr{B}}}(\theta)$ and $C_{X\to Y}(\hat{\theta})$ the lowest complexity score corresponding to the model $\mathscr{B}_{\widehat{\mathscr{G}},\hat{f},Q_N}$ with $\widehat{\mathscr{G}} = X \to Y$.

  If $C_{X\to Y}(\hat{\theta}) < C_{Y\to X}(\hat{\theta})$, it is decided that $X \to Y$, otherwise it is decided that $Y \to X$.
- Methods such as those by Stegle et al. [31] that compute a weighted bi-criteria aggregation between fit and complexity $\mathscr{A}_{\widehat{\mathscr{B}}}(\theta) = S_{\widehat{\mathscr{B}}}(\theta) + \lambda C_{\widehat{\mathscr{B}}}(\theta)$. We use the same notation $\hat{\theta}$ for the set of parameters that minimizes the score $\mathscr{A}_{\widehat{\mathscr{B}}}(\theta)$. If $\mathscr{A}_{X\to Y}(\hat{\theta}) < \mathscr{A}_{Y\to X}(\hat{\theta})$, it is decided that $X \to Y$, otherwise it is decided that $Y \to X$.

  This bi-objective aggregation corresponds for example to cause-effect inference based on the MML principle. In this case, the total score is directly $S_{\widehat{\mathscr{B}}}(\theta) + C_{\widehat{\mathscr{B}}}(\theta)$, without any aggregation parameter λ. However this aggregation parameter is "implicitly" set in the chosen prior $\pi(\theta)$.

With all these different approaches, a *causal score* that measures the confidence of the approach in the causal orientation can be defined as the difference between the two scores evaluated in each direction. For example for the methods that reason at fixed complexity and compare the fit scores if $S_{X\to Y}(\hat{\theta}) < S_{Y\to X}(\hat{\theta})$, $X \to Y$ is preferred with causal score $\Delta_{X\to Y} = S_{Y\to X}(\hat{\theta}) - S_{X\to Y}(\hat{\theta}) > 0$.

3.5 Review of the Main Pairwise Methods

According to the complexity/fit trade-off framework defined in the previous section we can propose a reading grid for the main algorithms developed in the literature (Table 3.1).

We introduce a taxonomy of methods according to the type of functional involved, their fit scores, their complexity scores and their way to deal with the trade-off fit/complexity.

Three main families of methods emerge:

1. Methods with a restricted class of mechanisms.
2. Non-parametric methods.
3. Methods exploiting the independence between cause and mechanism.

We present with more details these families in the next Sects. 3.5.1–3.5.3.

Table 3.1 Classification of the main algorithms for cause-effect inference proposed in the literature

Method	Reference	Interaction noise/cause	Mechanism	Fit score	Complexity score	Trade-off F/C
I) Restricted class						
Lingam	[28]	Additive noise	Linear regression	Log-likelihood	Restricted class	Compare fit scores
ANM-HSIC	[12]	Additive noise	Gaussian process	Indep. noise/cause	Restricted class	Compare fit scores
PNL	[34]	Post additive noise	Neural network	Indep. noise/cause	Restricted class	Compare fit scores
Markov kernel	[32]	Mixed noise	Markov kernel	Log-likelihood	Restricted class	Compare fit scores
II) Non-parametric						
GPI-MML	[31]	Mixed noise	Gaussian process	Log-likelihood	MML	Aggregation
CGAN-C2ST	[16]	Mixed noise	Neural networks	GAN	Flexible class	Compare fit scores
CGNN	[7]	Mixed noise	Neural networks	MMD	Flexible class	Compare fit scores
III) Indep. cause/mech						
IGCI	[4]	Deterministic function	Invertible function	–	Indep. cause/mech.	Compare complexities
CURE	[27]	Mixed noise	Gaussian process	Log-likelihood	Indep. cause/mech.	Compare complexities
CRACK	[17]	Mixed noise	–	MDL	Indep. cause/mech.	Aggregation
KCDC	[19]	Mixed noise	Kernel embedding	–	Indep. cause/mech.	Compare complexities
RECI	[2]	Mixed noise	Conditional expectation	–	Indep. cause/mech.	Compare complexities

3.5.1 Methods with a Restricted Class of Mechanisms

This first class of methods developed from 2006 to 2010 rely on restrictive class of admissible mechanisms $\mathscr{F}$ and focus on identifiability results. The main idea is to show that in some cases there exist $f \in \mathscr{F}$ and Q_N such that the hypothesis $\mathscr{B}_{\mathscr{G},f,Q_N}$ holds in the causal direction $\mathscr{G}$ with respect to the observed data while there do not exist any $f' \in \mathscr{F}$ nor Q'_N such that $\mathscr{B}_{\mathscr{G}',f',Q'_N}$ holds in the opposite causal direction $\mathscr{G}'$.

These methods range from very simple class of functions with LINGAM [28] to more complex class of functions such as PNL [34], with each time a trade-off between the identifiability and the generality of the proposed approach as depicted on Fig. 3.15.

Indeed, when the class of function is very restricted, there are fewer non identifiable cases, but the model can only successfully be used when encountering very specific observed data (such as data generated by linear mechanisms). When the class of functions becomes larger, the model becomes more general and can be used for more types of distribution, but at the cost of more non identifiable cases. In the extreme case of a completely general model without restriction on the class of mechanisms, all pairs become non identifiable as shown in Sect. 3.3.

3.5.1.1 Pairwise LiNGAM Inference

The LiNGAM [28] method was first developed for directed acyclic graph orientation for more than two variables. LiNGAM handles linear structural equation models, where each variable is continuous and modeled as:

$$X_i := \sum_k \alpha_k P_a^k(X_i) + E_i, i \in [\![1, n]\!], \tag{3.26}$$

with $P_a^k(X_i)$ the k-th parent of X_i and α_k a real value. Assuming further that all probability distributions of source nodes in the causal graph are non-Gaussian, [28] show that the causal structure is fully identifiable (all edges can be oriented).

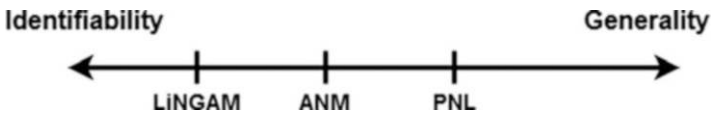

Fig. 3.15 Diagram on the identifiability/generality trade-off

Model

In the bivariate case, the authors assume that the variables X and Y are non Gaussian, as well as standardized to zero mean and unit variance. The goal is to distinguish between candidate linear causal models.

The first is denoted by $X \to Y$ and defined as:

$$Y := \rho X + N_Y \text{ with } X \perp\!\!\!\perp N_Y. \tag{3.27}$$

The second model with orientation $Y \to X$ is defined as:

$$X := \rho Y + N_X \text{ with } Y \perp\!\!\!\perp N_X. \tag{3.28}$$

The parameter ρ is the same in the two models because it is equal to the correlation coefficient between X and Y.

Identifiability Result

A theoretical identifiability has been derived by Shimizu et al. [28] who prove that if $P_{X,Y}$ admits the linear model $Y := aX + N_Y$ with $X \perp\!\!\!\perp N_Y$ (model 1), then there exist $b \in \mathbb{R}$ and a random variable N_X such that $X := bY + N_X$ with $Y \perp\!\!\!\perp N_X$ (model 2) if and only if X and N_Y are Gaussian.

In different words there is only one non-identifiable case corresponding to the linear Gaussian case presented in Sect. 3.3.1. Moreover if X or N_Y is non-Gaussian, when the candidate linear model with orientation $X \to Y$ holds, the candidate linear model with orientation $Y \to X$ does not hold.[5]

Practical Evaluation

The candidate models correspond to a restrictive class of mechanism and the comparison of the candidate models is based on comparison of fit scores at fixed complexity. The fit score used for comparison is the likelihood score as defined in Sect. 3.4.2.1 and derived by Hyvärinen and Smith [13] in this case as:

$$S_{X\to Y} = -\left[\sum_{i=1}^{n} G_X(x^i) + G_{N_Y}\left(\frac{y^i - \rho x^i}{\sqrt{1-\rho^2}}\right)\right] + n\log(1-\rho^2), \tag{3.29}$$

[5]However as discuss in Sect. 3.3.2 there always exists a potential nonlinear model $Y \to X$ that holds ($X := f_Y(Y, N_X)$) but this model is assumed to be more complex and is rejected due to the prior assumption that linear mechanisms are simpler.

where $G_X(u) = \log P_X(u)$ and G_{N_Y} is the standardized log probability distribution function of the residual when regressing Y on X. $S_{Y\to X}$ is computed similarly and the causal direction is decided by comparing $S_{X\to Y}$ with $S_{Y\to X}$.

3.5.1.2 Additive Noise Model

An extension of the previous model to deal with nonlinear mechanism has been derived by Hoyer et al. [12].

Model

A bivariate additive noise model (ANM) $X \to Y$ is defined as:

$$Y := f_Y(X) + N_Y \text{ with } X \perp\!\!\!\perp N_Y. \tag{3.30}$$

$f_Y : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function.

Identifiability Result

Hoyer et al. [12] proved that the ANM model is generally identifiable, saying that if $P_{X,Y}$ satisfies an additive noise model with orientation $X \to Y$, $P_{X,Y}$ cannot satisfy an additive model with orientation $Y \to X$.

However, specific cases have been identified where ANM is non-identifiable (see [35]). In particular the linear Gaussian case presented in Sect. 3.3.1 is non identifiable.

Practical Evaluation Based on Independence Score (ANM-HSIC)

In the original paper of [12] the ANM is seen as a discriminative model and the fit score is based on an independence test between noise and cause (Sect. 3.4.2.2). More precisely for the two alternatives $X \to Y$ and $Y \to X$, the estimated mechanisms $\hat{f}_Y$ and $\hat{f}_X$ are obtained via Gaussian process regressions. These estimated regression functions are used to estimate the residuals $\hat{n}_Y = y - \hat{f}_Y(x)$ and $\hat{n}_X = x - \hat{f}_X(y)$. The scores $S_{X\to Y}$ and $S_{Y\to X}$ correspond respectively to kernel HSIC independence test [9] between $\hat{n}_Y$ and x (for $X \to Y$) and between $\hat{n}_X$ and y (for $Y \to X$).

If $S_{X\to Y} < S_{Y\to X}$ it is decided that $X \to Y$. If $S_{X\to Y} > S_{Y\to X}$ it is decided that $Y \to X$.

3.5.1.3 Post Nonlinear Model

The ANM may fail to recover the causal direction when the noise is not additive. Therefore a generalization of ANM called Post-NonLinear model (PNL) that takes into account nonlinear interactions between the cause and the noise has been proposed by Zhang and Hyvärinen [34].

Model

A bivariate Post-NonLinear Model (PNL) $X \to Y$ is defined as:

$$Y := \hat{g}_Y(\hat{f}_Y(X) + N_Y) \text{ with } X \perp\!\!\!\perp N_Y. \tag{3.31}$$

$\hat{f}_Y : \mathbb{R} \to \mathbb{R}$ and $\hat{g}_Y : \mathbb{R} \to \mathbb{R}$ are two Borel measurable functions. $\hat{g}_Y$ is assumed to be invertible.

One can notice that LiNGAM and ANM are special cases of PNL. Indeed by setting $\hat{g}_Y$ to the identity function we recover the ANM. By choosing also $\hat{f}_Y$ to be linear and one of X or N_Y to be non Gaussian we recover the LiNGAM model.

Identifiability Result

Zhang and Hyvärinen [35] proved that the PNL model is generally identifiable, saying that if $P_{X,Y}$ satisfies a PNL model $X \to Y$, $P_{X,Y}$ cannot satisfy a PNL $Y \to X$, except when specific conditions are encountered. The set of non-identifiable distribution $P_{X,Y}$ is larger than for ANM. However PNL is more general and can handle more types of observed distribution.

Practical Evaluation

The model has been evaluated in [34] using an independence score between cause and noise (cf. Sect. 3.4.2.2).

As $\hat{g}_Y$ is assumed to be invertible, the idea is that the noise variable in Eq. (3.31) can be recovered from $P_{X,Y}$ as:

$$\hat{N}_Y = g_Y^{-1}(Y) - f_Y(X). \tag{3.32}$$

The noise variable is then estimated by functions l_1 and l_2 such as $\hat{N}_Y = l_1(Y) - l_2(X)$ with $\hat{N}_Y$ independent of X. It comes back to solve a constrained nonlinear ICA problem, that can be achieved by minimizing $I(X, \hat{N}_Y; \theta)$, the mutual information between X and $\hat{N}_Y$ [34] with respect to the parameter of the model θ. Symmetrically, an optimization of $I(Y, \hat{N}_X; \theta)$ is performed.

The causal direction $X \to Y$ is preferred if $I(X, \hat{N}_Y; \hat{\theta}) < I(Y, \hat{N}_X; \hat{\theta})$, $Y \to X$ otherwise.

3.5.1.4 Causal Inference by Choosing Graphs with Most Plausible Markov Kernel

In [32], the generative candidates model are derived from plausible Markov kernels and evaluated with log-likelihood scores.

Model

For a generative bivariate model $X \to Y$, the evaluation of cause and mechanism are the following:

- the distribution of the modeled cause Q_X is recovered by maximizing the entropy $H(Q_X)$ under constraints on the first and second moments of Q_X to make them correspond to those of the observed data P_X.
- following the same idea the distribution of the mechanism $Q_{Y|X}$ is modeled by maximizing the entropy $H(Q_{Y|X})$ under constraints on the first and second moments.

Evaluation

Once the model is recovered, log-likelihood scores of the model $Q_X Q_{Y|X}$ and $Q_Y Q_{X|Y}$ are computed as explained in Sect. 3.4.2.1 and the causal direction is determined by comparing the two scores.

3.5.2 Non-parametric Methods

The methods presented in the last section always assumed that the causal mechanisms belong to a restricted class of functions $\mathscr{F}$. However, this a priori restriction poses serious practical limitations when the task is to infer the causal direction on real data. Indeed in reality the mechanisms are often far from linearity and the interaction between noise and cause may be more complex than additive or even post-nonlinear noise. This is why more general methods have been proposed following pioneer works by Stegle et al. [31]. These methods in general offer better overall results on real data as they are more flexible, but they come with a loss of theoretical identifiability results, as no explicit restriction on the class of function is imposed (Sect. 3.3). The causal direction is often recovered by setting a smooth prior on the complexity of the mechanisms.

3.5.2.1 Probabilistic Latent Variable Models

A fully non-parametric Bayesian approach was proposed by Stegle et al. [31]. The name of the algorithm is GPI for Gaussian Process Inference.

Model

The approach aims to address the most general formulation of generative bivariate model with orientation $X \to Y$ with complex interaction between cause and noise: $Y := f_Y(X, N_Y)$.

The distribution of cause in GPI if modeled with a Gaussian mixture model with k modes:

$$Q(x^i|\theta_X) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(x^i|\mu_j, \sigma_j^2), \tag{3.33}$$

with parameters $\theta_X = \{(\alpha_j, \mu_j, \sigma_j)\}_{j=1}^k$.

The mechanism of GPI is a Gaussian process with zero mean and square exponential covariance function $\mathbf{K}_{\theta_Y}$ whose entry (i, j) is:

$$k_{\theta_Y}((x^i, n_Y^i), (x^j, n_Y^j)) = \lambda_Y^2 \exp\left(-\frac{(x^j - x^i)^2}{2\lambda_X^2}\right) \exp\left(-\frac{(n_Y^j - n_Y^i)^2}{2\lambda_N^2}\right), \tag{3.34}$$

where $\theta_Y = (\lambda_X, \lambda_Y, \lambda_N)$ are hyper parameters. Gamma priors are set on all these lambda parameters.

Practical Evaluation

The model is then evaluated using the MML framework exposed in Sect. 3.4.3.2. According to [31], for a GPI model in the direction $X \to Y$, the global aggregated MML score between fit and complexity can be expressed as:

$$\mathscr{A}_{X \to Y} = \mathscr{A}_X + \mathscr{A}_{Y|X}, \tag{3.35}$$

where the MML score for the cause is:

$$\mathscr{A}_X = \min_{\theta_X} \left(\sum_{j=1}^{k} \log(\frac{n\alpha_j}{12}) + \frac{k}{2}\log\frac{N}{12} + \frac{3k}{2} - \log Q(x|\theta_X) \right). \tag{3.36}$$

The MML score for the mechanism is:

$$\mathscr{A}_{Y|X} = \underset{\theta_Y, n_Y}{min} \Bigg(-\log\pi(\theta_Y) - \log\mathscr{N}(n_Y|0, \mathbf{I}) \\ -\log\mathscr{N}(y|0, \mathbf{K}_{\theta_Y}) + \sum_{i=1}^{n} \log|M_i \mathbf{K}_{\theta_Y}^{-1} y| \Bigg), \tag{3.37}$$

where the matrix M is defined for index (i, j) as $M_{i,j} = \frac{\delta k_{\theta_Y}}{\delta n_Y}((x^i, n_Y^i), (x^j, n_Y^j))$.

If $\mathscr{A}_{X\to Y} < \mathscr{A}_{Y\to X}$, it is decided that $X \to Y$, $Y \to X$ otherwise.

3.5.2.2 Conditional GAN for Causal Discovery

Lopez-Paz and Oquab [16] propose to use Conditional Generative Adversarial Networks, or CGANs [18] in order to model realistic causal mechanisms.

Model

The model with direction $X \to Y$ takes the form of a discriminative model, where the causal mechanism is defined by a one hidden unit neural network $\hat{f}_Y$ with parameter θ_Y. The variable Y is generated as:

$$Y := \hat{f}_Y(X, N_Y). \tag{3.38}$$

N_Y is independent from X and drawn according to standard normal distribution $\mathscr{N}(0, 1)$. X is the conditioning variable input to the generator and follow the observed distribution P_X.

Evaluation

In order to train this conditional GAN, a discriminative neural network d with parameter ω is used and the problem amounts to solve this min max optimization problem:

$$S_{X\to Y}(\hat{\theta}) = \underset{\theta}{min}\ \underset{\omega}{max}\ \Big(\mathbb{E}_{x,y}[\log d_\omega(x, y)] + \mathbb{E}_{x,n_Y}[\log(1 - d_\omega(x, \hat{f}_Y(x, n_Y; \theta)))]\Big). \tag{3.39}$$

Then the principle used for causal discovery is to learn two CGANs: one with a generator $\hat{f}_Y$ from X to Y to synthesize the dataset $\widehat{\mathscr{D}}_{X\to Y} = \{(x^i, \hat{y}^i)\}_{i=1}^n$, and the other with a generator $\hat{f}_X$ from Y to X to synthesize the dataset $\widehat{\mathscr{D}}_{Y\to X} = \{(\hat{x}^i, y^i)\}_{i=1}^n$. Then, the causal direction is decided to be $X \to Y$ if the two-sample test statistic between the real sample $\mathscr{D} = \{(x^i, y^i)\}_{i=1}^n$ and $\widehat{\mathscr{D}}_{X\to Y}$ is smaller than the one between $\mathscr{D}$ and $\widehat{\mathscr{D}}_{Y\to X}$.

3.5.2.3 Causal Generative Neural Network

Goudet et al. [7] proposed and extension of [16] for multivariate causal discovery called CGNN for Causal Generative Neural Network. As in [16] the mechanisms are modelled with generative neural networks.

If the joint density function h of $P_{X,Y}$ is continuous and strictly positive on a compact and convex subset of $\mathbb{R}^2$, and zero elsewhere, it has been shown that there exist two CGNNs $X \to Y$ and $Y \to X$, that approximates $P_{X,Y}$ with arbitrary accuracy. This result highlights the generality of the approach. However it raises also the issue that the CGNN can reproduce equally well the observational distribution in both directions. This non-identifiability issue is empirically mitigated by restricting the class of CGNNs considered, and specifically limiting the number n_h of hidden neurons in each causal mechanism. This parameter n_h can be seen as a complexity parameter that governs the CGNN ability to model the causal mechanisms: too small n_h, and data patterns may be missed; too large n_h, and overly complicated causal mechanisms may be retained.

Practical Evaluation

For practical use in the cause-effect pair setting n_h is empirically set to 30 hidden units in [7]. The fit score of the model is evaluated with a kernel two sample test between the sample $\mathscr{D}$ coming from observed distribution $P_{X,Y}$ and the sample $\widehat{\mathscr{D}}$ coming from the generated distribution $Q_{X,Y}$:

$$S_{X\to Y}(\theta) = \widehat{\mathrm{MMD}}_k(\mathscr{D}, \widehat{\mathscr{D}}), \tag{3.40}$$

where $\widehat{\mathrm{MMD}}_k(\mathscr{D}, \widehat{\mathscr{D}})$ is the empirical Maximum Mean Discrepancy (MMD) [10] defined as:

$$\widehat{\mathrm{MMD}}_k(\mathscr{D}, \widehat{\mathscr{D}}) = \frac{1}{n^2}\sum_{i,j=1}^{n} k(\mathbf{z}_i, \mathbf{z}_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\mathbf{z}_i, \hat{\mathbf{z}}_j), \tag{3.41}$$

where $\mathbf{z} = [x, y]$ is the two dimensional vector composed of x and y. The kernel k is usually taken as the Gaussian kernel ($k(\mathbf{z}, \mathbf{z}') = \exp(-\gamma \|\mathbf{z} - \mathbf{z}'\|_2^2)$).

For a fix number of hidden units n_h in each mechanism $\hat{f}_X$ and $\hat{f}_Y$, the causal direction is then based on the comparison of these best fit scores $S_{X\to Y}(\hat{\theta})$ and $S_{Y\to X}(\hat{\theta})$ in both directions.

3.5.3 Methods That Exploit Independence Between Cause and Mechanism

This class of methods exploits the notion of complexity of the causal mechanisms from a different perspective. They are based on the principle stated in Sect. 3.4.3.2 saying that when $X \rightarrow Y$ the distribution of the cause P_X should not contain information that can be useful to derive the conditional model $Q_{Y|X}$ on the data.

3.5.3.1 NonLinear Deterministic Mechanism

One of the first method that exploits the postulate of independence between cause and mechanism is the Information Geometric Causal Inference algorithm (IGCI) [4].

Model

When $X \rightarrow Y$, the authors assume that Y was generated from X by a nonlinear deterministic and invertible function:

$$Y := h(X). \tag{3.42}$$

Then the authors exploit a certain type of independence between P_X and the estimate of the function h. In particular, they interpret $x \rightarrow P_X(x)$ and $x \rightarrow \log h'(x)$ as random variables on the probability space $[0, 1]$. Then they compute the covariance of these two random variables with respect to the uniform distribution on $[0, 1]$:

$$\begin{aligned} \mathrm{Cov}(\log h', P_X) &= \mathbb{E}(\log h' \dot{P}_X) - \mathbb{E}(\log h')\mathbb{E}(P_X) \\ &= \int_0^1 \log h'(x) \dot{P}_X(x) dx - \int_0^1 \log h'(x) dx \int_0^1 P_X(x) dx. \end{aligned} \tag{3.43}$$

The authors show that if $\mathrm{Cov}(\log h', P_X)$ is close to zero, meaning that P_X does not contain information about $Q_{Y|X}$. Moreover it implies that for the reverse direction P_Y and $\log h'^{-1}$ are positively correlated. Therefore P_Y contains information about $Q_{X|Y}$, which implies an asymmetry between X and Y.

Practical Evaluation

To evaluate this model, the authors compute:

$$C_{X\to Y} = \int_0^1 \log h'(x) P_X(x) dx. \tag{3.44}$$

$$C_{Y\to X} = \int_0^1 \log h'^{-1}(y) P_Y(y) dy = -S_{X\to Y}. \tag{3.45}$$

The algorithm IGCI infers $X \to Y$ whenever $C_{X\to Y}$ is negative. The authors also show that the evaluation can be simplified into:

$$C_{X\to Y} = H(Y) - H(X). \tag{3.46}$$

Then it is decided that $X \to Y$ if $H(X) > H(Y)$ and $Y \to X$ otherwise.

3.5.3.2 Unsupervised Inverse Regression

Sgouritsa et al. [27] propose a method based on the idea that if $X \to Y$, P_X should not contain information about $P_{Y|X}$, while P_Y may contain information about $P_{X|Y}$. Therefore the estimation of $Q_{Y|X}$ based on P_X should be less accurate than the estimation of $Q_{X|Y}$ based on P_Y.

Model

The causal mechanism $Q_{Y|X}$ is modeled with a Gaussian process latent variable model whose likelihood function with respect to the data is given by:

$$Q(y|x, \theta_Y) = \mathcal{N}(y|0, \mathbf{K}(x) + \sigma_n^2 \mathbf{I})). \tag{3.47}$$

The entry (i, j) of $\mathbf{K}$ is:

$$k((x^i), (x^j)) = \sigma_f^2 \exp\left(-\frac{(x^j - x^i)^2}{\ell^2}\right), \tag{3.48}$$

with the set of parameters $\theta_Y = (\ell, \sigma_f, \sigma_n)$.

The idea is then to estimate $Q_{X|Y}^{unsup}$ based only on the sample $\mathbf{y}* = \{y^i\}_{i=1}^n$ of P_Y with unsupervised Gaussian process regression and to compare it with the estimate $Q_{X|Y}^{sup}$ based on the sample $\mathscr{D} = \{(x^i, y^i)\}_{i=1}^n$ of $P_{X,Y}$. Negative log-likelihoods scores of both models are then computed as $C_{X|Y}^{unsup} = -\frac{1}{n}\sum_{i=1}^n \log Q_{X|Y}^{unsup}(x^i|y^i)$ and $C_{X|Y}^{sup} = -\frac{1}{n}\sum_{i=1}^n \log Q_{X|Y}^{sup}(x^i|y^i)$. The accuracy of the estimation of $Q_{X|Y}$ based only on P_Y is then accessed by computing $C_{X|Y} = C_{X|Y}^{unsup} - C_{X|Y}^{sup}$. A symmetric evaluation is done for $C_{Y|X}$. The causal direction $X \to Y$ is then preferred to $Y \to X$ if $C_{X|Y} < C_{Y|X}$.

3.5.3.3 Causal Inference via Kernel Deviance Measures

Mitrovic et al. [19] exploit the same postulate that $Q_{Y|X}$ should be independent of P_X whenever $X \to Y$. The idea is that the estimate of the conditional distribution $\{Q_{Y|X=x^i}\}_{i=1}^n$ should be less sensitive to the different values x^i taken by the variable X than the conditional $\{Q_{X|Y=y^i}\}_{i=1}^n$ is sensitive to the different values of $Y = y^i$.

Model

For $i = 1..n$, $\{Q_{Y|X=x^i}\}_{i=1}^n$ is evaluated with a conditional RBF kernel embedding into the Hilbert space of infinitely differentiable functions.

Evaluation

A score for the causal direction $X \to Y$ is evaluated as the deviance of this set of conditional embeddings with respect to the Reproducing Kernel Hilbert Space (RKHS) norm as:

$$C_{X\to Y} = \frac{1}{n}\sum_{i=1}^{n}\left(|| \mu_{Y|X=x^i} ||_{\mathscr{H}_{\mathscr{Y}}} - \frac{1}{n}\sum_{j=1}^{n} || \mu_{Y|X=x^j} ||_{\mathscr{H}_{\mathscr{Y}}} \right)^2 . \tag{3.49}$$

$C_{Y\to X}$ is computed analogously and the causal direction $X \to Y$ is preferred if $C_{X\to Y} < C_{Y\to X}$.

3.5.3.4 Cause-Effect Inference by Comparing Regression Errors

A new method proposed by Bloebaum et al. [2] called RECI is based on a slightly different idea than the assumption of independence between cause and mechanism. The authors assume that the causal mechanism represents *a law of nature* that persists when the distribution of the cause and the distribution of the noise "change due to changing background conditions".

Model

They denote by ϕ the function that minimizes the expected least-squares error when predicting Y from X, $\phi(x) = \mathbb{E}[Y|X = x]$, and ψ the minimizer of the least-squares error for predicting X from Y, $\phi(y) = \mathbb{E}[X|Y = y]$.

For a model with the causal orientation $X \to Y$, the authors rewrite the bivariate functional causal model of Definition 3.1 under the following new form:

$$Y_\alpha := \phi(X) + \alpha N, \tag{3.50}$$

where $\alpha \in \mathbb{R}^+$ and N is a noise variable not necessarily independent of X.

They show under the regime of almost deterministic relations (when $\alpha \to 0$) that it implies:

$$\mathbb{E}\left[(Y - \phi(X))^2\right] \leq \mathbb{E}\left[(X - \psi(Y))^2\right], \tag{3.51}$$

which can be translated into "the MSE of regressing Y on X is lower that the MSE of regressing X on Y".

To obtain this result, the authors make the main following assumptions that we briefly summarize here:

- **Invertible function**: ϕ is a strictly monotonically increasing (or decreasing) twice differentiable function.
- **Compact supports**: the distribution of X has compact support.
- **Independence postulate**: the functions $x \mapsto \phi'(x)$ and $x \mapsto \text{Var}[N|X = x]$ $p_X(x)$ that define random variables are assumed to be uncorrelated, which is formally stated after re-scaling of the variables between 0 and 1 as:

$$\int_0^1 \phi'(x)\text{Var}[N|X{=}x]\, p_X(x)dx - \int_0^1 \phi'(x)dx \int_0^1 \text{Var}[N|X{=}x]\, p_X(x)dx{=}0. \tag{3.52}$$

Practical Evaluation

The method RECI consists in fitting a non-linear regression model on both directions after re-scaling between 0 and 1 both variables and comparing the mean squared error losses. The regression models used by the authors may be logistic functions, polynomial functions or neural networks. In practice the authors report better empirical results with simple polynomial regression models such as the shifted monomial functions $ax^3 + c$.

3.6 Experimental Comparison of Cause-Effect Inference Algorithms

In the last section we have presented an overview of the main methods proposed for the cause-effect pair problem. Now we propose to evaluate the different algorithms whose code are available online on several cause-effect pair datasets.

3.6.1 Datasets

Five datasets with continuous variables are considered [6]:

- *CE-Cha*: 300 continuous variable pairs from the cause-effect pair challenge [11] that will be presented with more details in the next chapter. Here we only consider pairs with label $+1$ ($X \to Y$) and -1 ($Y \to X$) (notably the confounding case is excluded).
- *CE-Net*: 300 artificial pairs generated with a neural network initialized with random weights and random distribution for the cause (exponential, gamma, lognormal, laplace...).
- *CE-Gauss*: 300 artificial pairs without confounder sampled with the generator of [20]: $Y := f_Y(X, N_Y)$ and $X := f_X(N_X)$ with $N_X \sim P_{N_X}$ and $N_Y \sim P_{N_Y}$. P_{N_X} and P_{N_Y} are randomly generated Gaussian mixture distributions. Causal mechanisms f_X and f_Y are randomly generated Gaussian processes.
- *CE-Multi*: 300 artificial pairs generated with linear and polynomial mechanisms. The effect variables are built with post additive noise setting ($Y := f_Y(X) + N_Y$), post multiplicative noise ($Y := f_Y(X) \times N_Y$), pre-additive noise ($Y := f_Y(X + N_Y)$) or pre-multiplicative noise ($Y := f_Y(X \times N_Y)$).
- *CE-Tueb*: 99 real-world cause-effect pairs from the *Tuebingen cause-effect pairs* dataset, version August 2016 [20]. This version of this dataset is taken from various domains: climate, census, medicine data.

For all variable pairs, the size n of the data sample is set with a maximum of 1500 for the sake of an acceptable overall computational load. To provide an overview of the type of pairs encountered in each dataset, one hundred pairs of each of them are displayed in Fig. 3.16 (the real pair with altitude and temperature given as introductory example corresponds to the first pair in the top left corner of *CE-Tueb*).

3.6.2 Algorithms

We compare the performance of the following algorithms presented in this chapter:

- **Best mse**: the method presented in the introduction that consists in fitting a non-linear Gaussian process regression model on both directions after re-scaling with zero mean and unit variance and comparing the mean squared error loss (mse). The direction corresponding to the fit with the lowest mse is preferred.

[6] The first four datasets are available at http://dx.doi.org/10.7910/DVN/3757KX. The *Tuebingen cause-effect pairs* dataset with real pairs is available at https://webdav.tuebingen.mpg.de/cause-effect/.

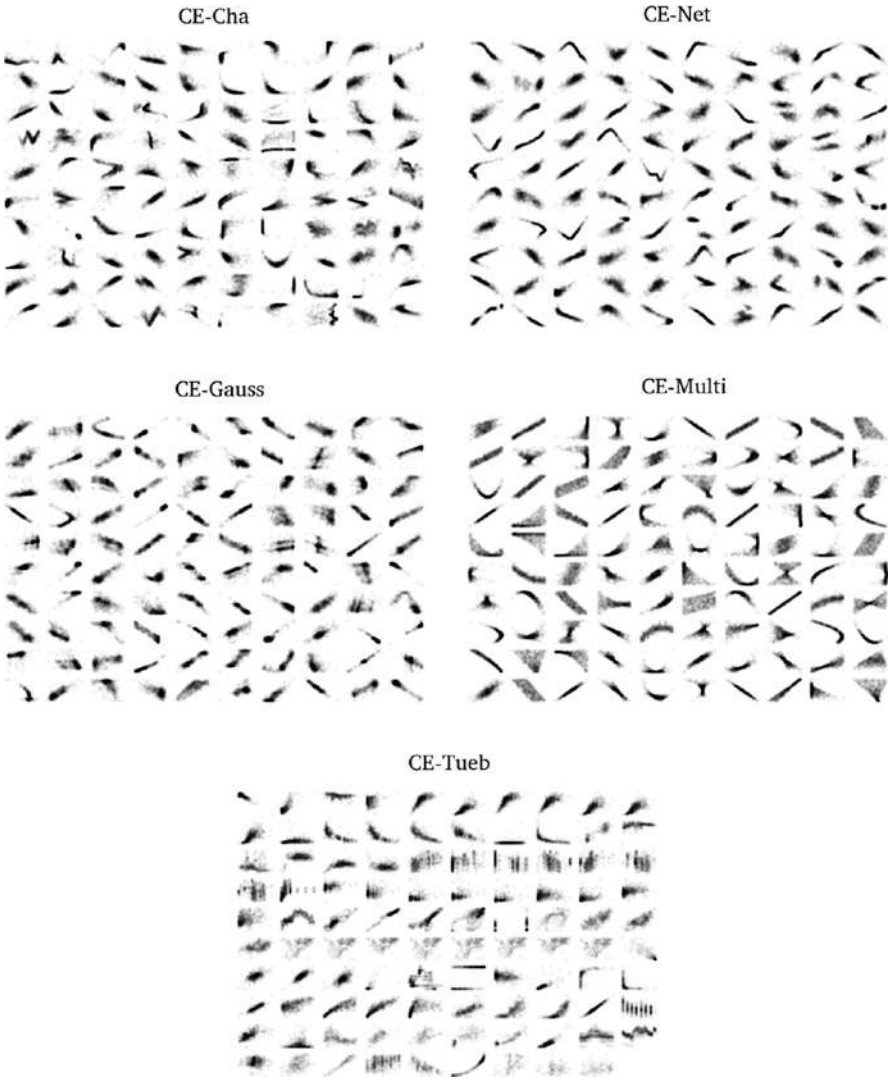

Fig. 3.16 (X,Y) plots of one hundred pairs of each dataset

- **RECI**: the practical implementation of this method is equivalent to the **Best mse** method except that the re-scaling is made between 0 and 1 and Gaussian process regression is replaced by polynomial regression with monomial function $ax^3 + c$ [2] (see Sect. 3.5.3.4).
- **LiNGAM**: a pairwise version of the method developed by Shimizu et al. [28] relying on Independent Component Analysis to identify the linear relations between variables (see Sect. 3.5.1.1).

- **ANM**: the Additive Noise Model [20] with Gaussian process regression and HSIC independence test of the residual (see Sect. 3.5.1.2).
- **IGCI**: the Information Geometric Causal Inference algorithm [4] with entropy estimator and Gaussian reference measure (see Sect. 3.5.3.1).
- **PNL**: the Post-NonLinear model with HSIC independence test [35] (see Sect. 3.5.1.3).
- **GPI**: the Gaussian Process Inference algorithm [31] based on the Minimum message length principle (see Sect. 3.5.2.1).
- **CGNN**: the Causal Generative Neural method [7] using neural networks to model causal mechanisms and Maximum Mean Discrepancy as fit score (see Sect. 3.5.2.3).

For the methods **Best mse** and **RECI** we use the implementation provided in the CausalDiscoveryToolbox.[7] For the implementation of the algorithms **ANM**, **IGCI**, **PNL**, **GPI** and **LiNGAM** we use the R program available at https://github.com/ssamot/causality. For **CGNN** we use the code available on github at https://github.com/GoudetOlivier/CGNN. We use default authors parameters for each algorithm implementation.

3.6.3 Performance Metric

The task of orienting each pair observed pair as $X \rightarrow Y$ or $Y \rightarrow X$ is a binary classification problem. We propose to use two scores to evaluate the performance of the algorithms:

- The **accuracy score** computed as the ratio of well oriented edges over the total number of edges for the datasets *CE-Cha*, *CE-Net*, *CE-Gauss* and *CE-Multi*. For the *CE-Tueb* dataset, a weighted accuracy is computed as in [20] in order to take into account dependent pairs from the same domain.
- The **area under the ROC curve** computed using the causal score given by each model that measures the confidence of the approach in the causal orientation for each pair. This score is defined as the difference between the two scores of each candidate model evaluated in both directions (cf. Sect. 3.4).

[7] Available online at https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox.

3.6.4 Results

Table 3.2 reports the area under the roc curve (AUROC) and the accuracy (in parenthesis) for each benchmark and each algorithm. The corresponding ROC curves are displayed in Fig. 3.17.

The method **LiNGAM** is outperformed as it uses linear mechanisms to model the data generative process which is in many cases unrealistic. **IGCI** does not seem to be very robust: it takes advantage of some specific features of the dataset, (e.g. the cause entropy being lower than the effect entropy in *CE-Multi*), but remains near chance level otherwise. The method **ANM** yields good results when the additive assumption holds (e.g. on *CE-Gauss*), but fails otherwise. **PNL**, less restrictive than **ANM**, yields overall good results compared to the former methods. The method **RECI** based on comparison of mean square error scores after proper re-scaling provides overall good results too. However it fails for more complex pairs due to restrictive assumptions on the causal mechanisms involved. Lastly, methods like **GPI** and **CGNN** that admit the most flexible class of causal mechanisms and non-parametric metrics to match the distributions perform well on most datasets, including the real-world cause-effect pairs *CE-Tueb*, in counterpart for a higher computational cost (resp. 32 min on CPU for GPI and 24 min on GPU for CGNN).[8]

Let us note that better scores for the dataset Tübingen (*CE-Tueb*) are reported in recent papers such as [16] (accuracy of 82.0%) and [19] (accuracy of 78.7%) using ensemble methods with the algorithms presented in Sects. 3.5.2.2 and 3.5.3.3.

Table 3.2 Area under the Precision Recall curve and accuracy in parenthesis on the five datasets and for all the algorithms

Methods	*CE-Cha*	*CE-Net*	*CE-Gauss*	*CE-Multi*	*CE-Tueb*	All
Best mse	50.0 (46.7)	86.4 (76.7)	18.7 (23.7)	46.9 (36.3)	61.3 (61.7)	61.3 (61.7)
RECI	59.0 (56.0)	66.0 (60.3)	71.0 (64.3)	94.7 (85.3)	70.5 (70.8)	73.8 (66.7)
LiNGAM	57.8 (55.7)	3.3 (36.7)	72.2 (77.3)	62.3 (63.3)	31.1 (44.3)	54.8 (57.8)
IGCI	55.6 (55.0)	57.4 (57.0)	16.0 (21.3)	77.8 (68.0)	63.1 (62.6)	52.8 (50.7)
ANM	43.7 (46.0)	87.8 (78.0)	90.7 (83.3)	25.5 (38.0)	63.9 (62.7)	60.7 (61.4)
PNL	78.6 (76.0)	75.6 (65.3)	84.7 (78.3)	51.7 (56.3)	73.8 (66.2)	71.9 (68.9)
GPI	71.5 (67.0)	88.1 (79.0)	90.2 (82.0)	73.8 (77.7)	70.6 (62.5)	79.9 (76.0)
CGNN	76.2 (71.6)	86.3 (75.3)	89.3 (81.0)	94.7 (87.3)	76.6 (75.9)	**86.5 (78.8)**

Underline values corresponds to best score for each dataset. *Bold* value corresponds to best overall score on all dataset.

[8]Computational times are measured on Intel Xeon 2.7Ghz (CPU) or on Nvidia GTX 1080Ti graphics card (GPU).

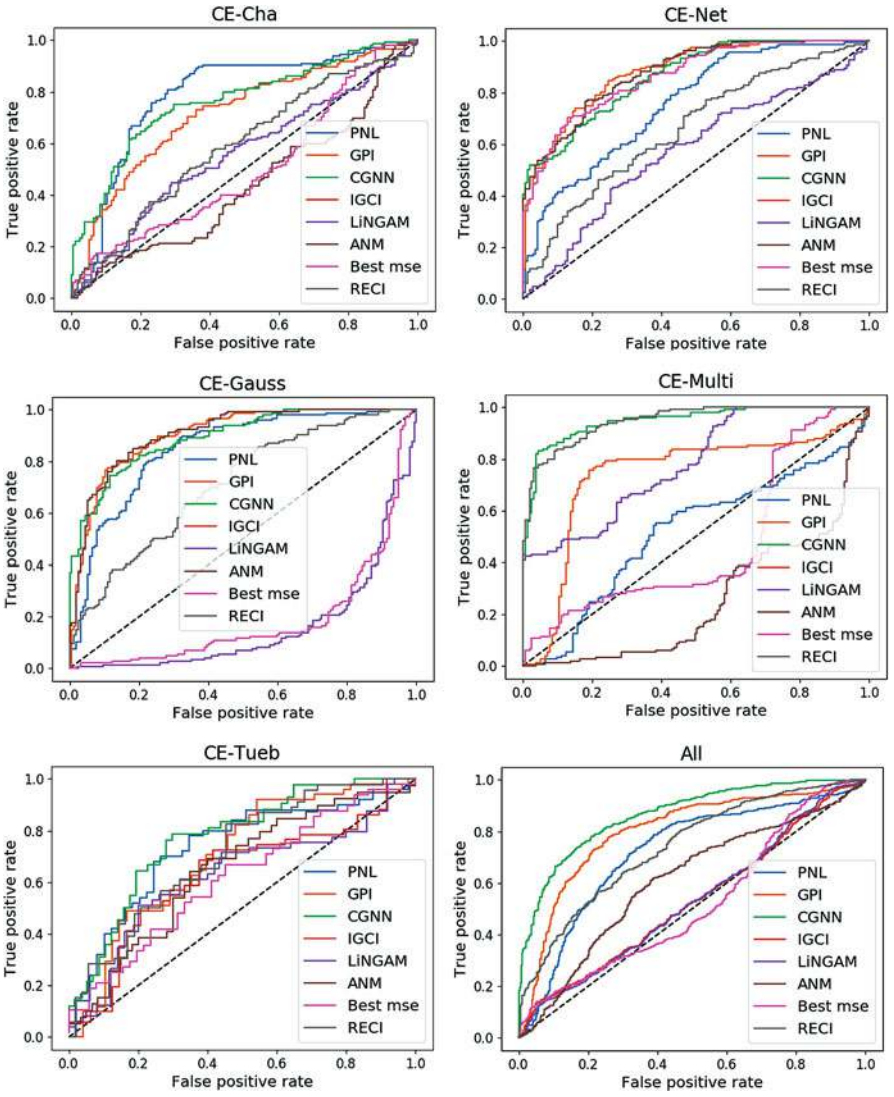

Fig. 3.17 ROC curves for the algorithms PNL, GPI, CGNN, IGCI, LiNGAM, ANM and Best mse on the different datasets. The last plot corresponds to the overall score on the five datasets

3.7 Discussion and Open Problems

The cause-effect inference problem is relatively new in the Machine Learning literature and there are still a lot of open problems to be addressed in order to build a robust tool that can be used by the practitioner to confirm for example if a given

treatment has an impact or not on a given disease or to discover if a gene has a regulatory power on an other one.

3.7.1 Relax the Causal Sufficiency Assumption

To build such useful tools for practitioners, one of the first assumption that needs to be relaxed is the causal sufficiency assumption as in a lot of real problems it is very rare to find a cause-effect problem that is not affected by hidden common confounder that can affect both variables such as age or gender.

One idea proposed in the literature is to model confounders by introducing correlation between the noise variables N_X and N_Y that affect X and Y as in [26] or by modelling all the unobserved confounding effects by a new noise variable N_{XY} entering in the generation process of X and Y [7, 15].

If we reformulate the generative bivariate framework of Sect. 3.4.1.1 in presence of confounders three alternative candidate models must be considered:

1. If $\mathscr{G} = X \rightarrow Y$:
   - $X := \hat{f}_X(N_X, N_{XY})$ with $N_X \sim Q_{N_X}$, $N_{XY} \sim Q_{N_{XY}}$ and $N_X \perp\!\!\!\perp N_{XY}$
   - then $Y := \hat{f}_Y(X, N_Y, N_{XY})$ with $N_Y \sim Q_{N_Y}$ and $N_Y \perp\!\!\!\perp N_{XY}$
2. If $\mathscr{G} = Y \rightarrow X$:
   - $Y := \hat{f}_Y(N_Y, N_{XY})$ with $N_Y \sim Q_{N_Y}$, $N_{XY} \sim Q_{N_{XY}}$ and $N_Y \perp\!\!\!\perp N_{XY}$
   - then $X := \hat{f}_X(Y, N_X, N_{XY})$ with $N_X \sim Q_{N_X}$ and $N_X \perp\!\!\!\perp N_{XY}$
3. If $\mathscr{G} = Y \leftrightarrow X$,
   - $X := \hat{f}_X(N_X, N_{XY})$ with $N_X \sim Q_{N_X}$, $N_{XY} \sim Q_{N_{XY}}$ and $N_X \perp\!\!\!\perp N_{XY}$
   - $Y := \hat{f}_Y(N_Y, N_{XY})$ with $N_Y \sim Q_{N_Y}$ and $N_Y \perp\!\!\!\perp N_{XY}$

A diagram of the model is presented in Fig. 3.18. We use the same notation for the vector of parameters as in Sect. 3.4.1.1 but with a new set of parameters $\theta_{N_{XY}}$ for the distribution of the common noise variable N_{XY}.

The same approach with a trade-off between fit and complexity could be used to compare these three alternative models in presence of potential confounders. An additional penalty term depending on the total number of edges $|\widehat{G}|$ in the model could be introduced like in score based methods for multivariate inference [3], in order to remove the eventual spurious link due to the confounding effect between X and Y. We have indeed $|\widehat{G}| = 1$ for the models with structures $X \rightarrow Y$ or $Y \rightarrow X$, and $|\widehat{G}| = 0$ for the models with structure $Y \leftrightarrow X$.

This framework could also take into account known variables as confounders. In this case the variable N_{XY} would be replaced by a set of observed variables (e.g. the latitude in the introductory example of this chapter with altitude and temperature).

Furthermore one can notice that this extension of the generative approach in presence of the confounding effect would not be possible for discriminative

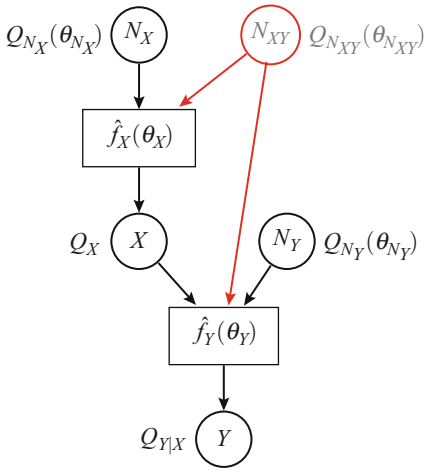

Fig. 3.18 Candidate generative models $X \to Y$ with a potential confounding effect modelled with the noise variable N_{XY}

approaches where the source cause is not modelled (when only comparing $P_{Y|X}$ with $P_{X|Y}$).

3.7.2 Need for Real Datasets of a Big Size

Another important problem that is often overlooked in the "cause-effect pair community" concerns benchmarking. As of the writing of this chapter, the main real data benchmark used to compare the different methods by the community is the Tübingen Dataset,[9] composed of only 100 hundred pairs which are often very similar. It is indeed very difficult to collect cause-effect pairs with enough data points from the real world with an authenticated known ground truth. But one must to keep in mind that it is very easy for most of the methods presented in this chapter to tune their hyper parameters (even unintentionally) in order to obtain the best results. This overfitting problem is often compounded by the fact that this dataset is, most of the time, not separated into train/validation/test sets. To overcome this problem a Cause-effect Pair Challenge has been proposed by Guyon [11] with real and artificial data generated with various mechanisms. It will be discussed in the next chapter.

[9] https://webdav.tuebingen.mpg.de/cause-effect/.

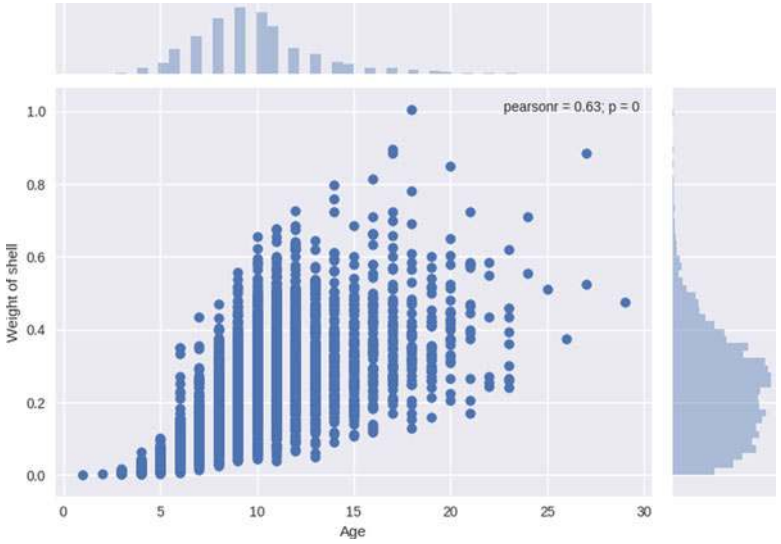

Fig. 3.19 Age of a snail (X-axis) and corresponding weight of its shell (Y-axis)

3.7.3 Biased Assessment Due to Artifacts in Data

It is often observed that the cause variable is discrete ordinal while the other is continuous. This induces an artificial asymmetry between cause and effect which could lead to biased assessments.

Let us consider for example real abalone data[10] representing the age of snails (X-axis) and the corresponding weight of its shell (Y-axis) as depicted in Fig. 3.19. The ground truth is *age* causes *weight of the shell*.

We see that due to the experimental conditions when collecting the data, the variable age is discrete (categorical ordered variable or ordinal variable) and not continuous. Because of this artifact on the age variable, the conditional $P_{\text{weight|age}}$ has really more chance to be simpler than $P_{\text{age|weight}}$. Therefore it favors approaches that compare only $P_{Y|X}$ with $P_{X|Y}$ and it may lead to inconsistent methods for the cause-effect pair problem as stated in Chap. 1 of this book.

An open problem could be to find a way to correct this bias when one encounters this type of data. This was done in the design of the dataset of the Cause-effect Pair Challenge.

[10] https://archive.ics.uci.edu/ml/datasets/Abalone.

3.7.4 Extension of the Generative Approach for Categorical Variables

In this chapter we have discussed the cause-effect pair problem for continuous variables, but the same idea could be used for categorical variables or mixed variables. To the best of our knowledge, only few attempts have been made to solve the cause-effect pair problem with categorical data [17, 24].

It may be explained by the fact that the cause-effect problem for categorical data may be really harder than for continuous data, because there is in general less information to exploit in the distribution of the noise and in the asymmetry of the causal mechanisms.

If one considers for example the extreme case of two binary variables, what kind of causal information can be really exploited from a matrix of size (2,2) that represents the repartition of the data points into the 4 different cases (1,0), (0,1), (1,1) and (0,0) ?

3.7.5 Extension of the Pairwise Setting for Complete Graph Inference

There is a need for methods that can really cross the bridge between the cause-effect pair problem and the complete problem of causal discovery with more than two variables. A way of research could be to propose an efficient approach for the general multivariate case that can potentially exploit all the information available, including the asymmetry between cause and effect and the conditional independence relations.

In this direction, an extension of the bivariate Post-NonLinear model (PNL) has been proposed [35], where an FCM is trained for any plausible causal structure, and each model is tested a posteriori for the required independence between errors and causes, but the main limitation is its super-exponential cost with the number of variables [35]. Another hybrid approach uses a constraint based algorithm to identify a Markov equivalence class, and thereafter uses bivariate modeling to orient the remaining edges [35]. For example, the constraint-based PC algorithm [30] can identify the v-structure, enabling the bivariate PNL method to further infer the remaining arrows that are not identifiable with conditional independence tests. However an effective combination of constraint-based and bivariate approaches requires a final verification phase to test the consistency between the v-structures and the edge orientations based on asymmetry.

An extension of the ANM model (Sect. 3.5.1.2) called CAM [25] proposed a consistent framework that can exploit the interaction and the asymmetry of the cause and effect, but the approach is restricted to the assumption of causal additive noise. The CGNN algorithm presented in Sect. 3.5.2.3 can be extended in the multivariate case [7]. It is more general than the CAM algorithm, but needs to start

from a supposed known skeleton of the causal graph and employ simple exploratory heuristic to explore the space of DAG due to computational reasons.

3.7.6 Computational Complexity Limitations

Some methods suffer from computational complexity limitations. In the bivariate case, there are only two alternative DAGs to compare, but for the multivariate case with more than 1000 variables, the number of different DAGs to consider grows exponentially. Therefore it is a real challenge to make the successful methods of the cause-effect problem scale for big data problem. In particular the methods that can model complex interactions between cause and noise, such as those using Gaussian process regressions or neural networks are often really slow to compute and do not scale well in term of number of variables and number of data points.

3.7.7 Relax Restrictive Assumptions on Causal Mechanisms

Another open problem concerns the fact that all these methods rely on specific assumptions on the underlying data generative process. All of them can work well when these specific assumptions are encountered in the observed data. This is why a new Machine Learning approach, presented in the next chapter, has appeared in recent years. It is based on the idea to combine all the successful algorithms presented in this chapter into a single meta algorithm that could benefit from the advantages of each of them.

3.8 Conclusion

We have briefly explained the difference between explanatory models and predictive models and seen that causal discovery consists in finding the best explanatory model of the data. We have then defined the problem of cause-effect inference in the bivariate setting and the main assumptions usually involved in the literature. Under these assumptions we have formalized the notion of bivariate structural equation model. We have then seen that the task of recovering the good causal direction from $X \rightarrow Y$ or $Y \rightarrow X$ consists in comparing alternative candidate models estimated from data. This comparison is always based in one way or another on a complexity/fit trade-off. We have reviewed how the complexity and the fit terms are usually evaluated in order to compute a causal score in both directions. It has led us to propose a reading grid of the main methods proposed in the literature.

Three main families have emerged: (1) methods that restrict the class of admissible causal mechanisms and focus on identifiability results; (2) methods

that do not explicitly restrict the class of admissible mechanisms and focus on the generality of the approach at the expense of theoretical identifiability guaranties; and (3) lastly methods that exploit the postulate that the cause should be independent of the causal mechanism.

We have then compared the main methods of the literature whose code is available online. It appears that methods that allow for flexible causal mechanism and complex realistic interactions can obtain consistent scores on a wide range of cause-effect problems (artificial and real data). These results need however to be confirmed on real datasets of bigger size with known ground truth.

One fruitful way of research could be to extend the best generative approach for causal discovery presented in this chapter to deal with potential confounding effects. Another interesting way of research could be to provide a theoretical framework that can unify the cause-effect pair methods presented in this chapter with the multivariate methods for causal discovery using conditional independence tests.

Software Packages

- R package including ANM, IGCI, PNL, GPI and LiNGAM algorithms available at the URL:
  https://github.com/ssamot/causality.
- Python package including Best mse, RECI, CGNN, ANM, IGCI and LiNGAM algorithms available at the URL:
  https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox.

Acknowledgements The authors would like to thank Daniel Rolland for proofreading this document, as well as the reviewers for their constructive feedback.

References

1. Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.
2. Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pages 900–909, 2018.
3. David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
4. Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 143–150, Arlington, Virginia, United States, 2010. AUAI Press. ISBN 978-0-9749039-6-5. http://dl.acm.org/citation.cfm?id=3023549.3023566.
5. Bruce Edmonds and Scott Moss. From kiss to kids–an 'anti-simplistic' modelling approach. In *International workshop on multi-agent systems and agent-based simulation*, pages 130–144. Springer, 2004.
6. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

7. Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.
8. Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
9. Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.
10. Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, Alexander J Smola, et al. A kernel method for the two-sample-problem. 19:513, 2007.
11. Isabelle Guyon. Chalearn cause effect pairs challenge, 2013. http://www.causality.inf.ethz.ch/cause-effect.php.
12. Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Neural Information Processing Systems (NIPS)*, pages 689–696, 2009.
13. Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.
14. Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.
15. Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.
16. David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.
17. Alexander Marx and Jilles Vreeken. Causal inference on multivariate and mixed-type data. *arXiv preprint arXiv:1702.06385*, 2017.
18. Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
19. Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *arXiv preprint arXiv:1804.04622*, 2018.
20. Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.
21. Judea Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 19(675–685):46, 2003.
22. Judea Pearl. *Causality*. Cambridge university press, 2009.
23. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
24. Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.
25. Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.
26. Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.
27. Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *Artificial Intelligence and Statistics*, pages 847–855, 2015.
28. Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

29. Galit Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.
30. Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.
31. Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Neural Information Processing Systems (NIPS)*, pages 1687–1695, 2010.
32. Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *ISAIM*, 2006.
33. Chris S Wallace and Peter R Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 240–265, 1987.
34. Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*, pages 157–164. JMLR. org, 2008.
35. Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.
36. Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):13, 2016.