Behaviormetrika
Vol.41, No.1, 2014, 65–98

Invited paper

LINGAM: NON-GAUSSIAN METHODS FOR ESTIMATING CAUSAL STRUCTURES

Shohei Shimizu*

In many empirical sciences, the causal mechanisms underlying various phenomena need to be studied. Structural equation modeling is a general framework used for multivariate analysis, and provides a powerful method for studying causal mechanisms. However, in many cases, classical structural equation modeling is not capable of estimating the causal directions of variables. This is because it explicitly or implicitly assumes Gaussianity of data and typically utilizes only the covariance structure of data. In many applications, however, non-Gaussian data are often obtained, which means that more information may be contained in the data distribution than the covariance matrix is capable of containing. Thus, many new methods have recently been proposed for utilizing the non-Gaussian structure of data and estimating the causal directions of variables. In this paper, we provide an overview of such recent developments in causal inference, and focus in particular on the non-Gaussian methods known as LiNGAM.

1. Introduction

In many empirical sciences, the causal mechanisms underlying various natural phenomena and human social behavior are of interest and need to be studied. Conducting a controlled experiment with random assignment is an effective method for studying causal relationships; however, in many fields, including the social sciences (Bollen, 1989) and the life sciences (Smith, 2012; Bühlmann, 2013), performing randomized controlled experiments is often ethically impossible or too costly. Thus, it is necessary and important to develop computational methods for studying causal relations based on data that are obtained from sources other than randomized controlled experiments. Such computational methods are useful for developing hypotheses on causal relations and deciding on possible future experiments to obtain more solid evidence of estimated causal relations (Maathuis et al., 2010; Pe'er and Hacohen, 2011; Smith, 2012).

A major framework for causal inference (Pearl, 2000) may be based on a combination of the counterfactual model of causation (Neyman, 1923; Rubin, 1974) and structural equation modeling (Bollen, 1989). The counterfactual model describes causation in terms of the relationships between the variables involved: generally speaking, if the value of a variable is changed and that of some other variable also changes, the former is the cause and the latter is the effect. Structural equation models are mathematical models that can be used to represent data-generating processes. Using structural equation models, one can mathematically represent the cause-and-effect relationships

Key Words and Phrases: Causal inference, Causal structure learning, Estimation of causal directions, Structural equation models, non-Gaussianity

* The Institute of Scientific and Industrial Research, Osaka University, Mihogaoka 8–1, Ibaraki, Osaka 567–0047, Japan. E-mail: sshimizu@ar.sanken.osaka-u.ac.jp

that are defined by using the counterfactual model.

Structural equation modeling provides a general framework for multivariate analysis and offers a powerful means of studying causal relations (Bollen, 1989; Pearl, 2000). However, in many cases, classical structural equation modeling is not capable of estimating the causal directions of variables (Bollen, 1989; Spirtes et al., 1993; Pearl, 2000). A major reason for this disadvantage is that this method explicitly or implicitly assumes the Gaussianity of data, and typically utilizes only the covariance structures of data for estimating causal relations. However, in many applications, it is common for non-Gaussian data to be obtained (Micceri, 1989; Hyvärinen et al., 2001; Smith et al., 2011; Sogawa et al., 2011; Moneta et al., 2013), which means that more information can be contained in the data distribution than in the covariance matrix. Bentler (1983) proposed making use of non-Gaussianity of data for estimating structural equation models, although this had not been extensively studied until recently.

New methods have since been proposed for utilizing the non-Gaussian structure of data and thereby estimating the causal directions of variables when studying causality (Dodge and Rousson, 2001; Shimizu et al., 2006). These methods have, in turn, led to the development of many additional methods, including latent confounder methods (Hoyer et al., 2008b; Shimizu and Hyvärinen, 2008), time series methods (Hyvärinen et al., 2010), nonlinear methods (Hoyer et al., 2009; Zhang and Hyvärinen, 2009b; Tillman et al., 2010) and discrete variable methods (Peters et al., 2011a). These non-Gaussian methods have been applied to the data studied in many fields, including economics (Ferkingsta et al., 2011; Moneta et al., 2013), behavior genetics (Ozaki and Ando, 2009; Ozaki et al., 2011), psychology (Takahashi et al., 2012), environmental science (Niyogi et al., 2010), epidemiology (Rosenström et al., 2012), neuroscience (Smith et al., 2011) and biology (Statnikov et al., 2012).

In this paper, we provide an overview of such recent developments in causal inference. In Section 2 of this paper, we first briefly review the basics of causal inference, including the counterfactual model of causation and its mathematical representation, based on structural equation models. We then discuss recent developments in methods applied to estimating causal structures, focusing in particular on the non-Gaussian methods known as Linear Non-Gaussian Acyclic Models (LiNGAM). We explain the basic LiNGAM model in Section 3, its estimation methods in Section 4 and its extensions in Section 5. Methods that form part of the LiNGAM group are capable of estimating a much wider variety of causal structures than classical methods.

2. Basics of causal inference

In this section, we provide a brief overview of causal inference (Bollen, 1989; Spirtes et al., 1993; Pearl, 2000). For an in-depth discussion, refer to Pearl (2000).

2.1 Counterfactual model of causation

We begin by introducing the concept of individual-level causation (Neyman, 1923; Rubin, 1974). Suppose that an individual named Taro is a patient with a certain disease. We want to know if a particular medicine cures his disease. To this end, we compare the consequences of two actions: i) Having him take the medicine; and ii) Having him *not* take the medicine. Suppose Taro recovers after three days later if he takes the medicine, but does not recover if he does not. Then, we can say that his taking the medicine caused his recovery within three days. Therefore, in terms of Taro, if the value of a binary variable x (1: takes the medicine, 0: does not take the medicine) is changed from 0 to 1, and that of a second binary variable y (1: recovers, 0: does not recover) changes from 0 to 1, it means that Taro's taking the medicine is the cause of his recovery.

However, a problem arises in such a situation: it is not possible to observe both of these consequences. This is because, once we observe the consequence of Taro taking the medicine, we can never observe that of him not taking the medicine. The former consequence is factual, since he actually took the medicine, while the latter is counterfactual, since it contradicts the fact. It is therefore impossible to compare the two consequences and derive a causal conclusion based on the data of the individual Taro, and this is known as the fundamental problem of causal inference (Holland, 1986).

Next, we introduce the concept of population-level causation (Neyman, 1923; Rubin, 1974). Suppose that all the individuals in a population are suffering from a certain disease. We want to know if a particular medicine will cure the disease in this population. To determine this, we compare the consequences of two actions: i) Having all the individuals in the population take the medicine; and ii) Having all the individuals *not* take the medicine. Suppose that the number of individuals who took the medicine and had recovered three days later is significantly larger than that of the individuals who did *not* take the medicine and recovered three days. Then, we can say that taking the medicine caused recovery in three days in this population.

Here, we encounter a similar problem as that in individual-level causation. That is, once we observe the consequence of all the individuals actually taking the medicine, we can never observe the consequence of them not taking the medicine. However, although individual-level causation generally cannot be determined, fortunately, it is sometimes possible to determine population-level causation, as discussed below.

2.2 Structural equation models for describing data-generating processes

In this subsection, we discuss structural equation models (SEMs) as a mathematical tool for describing the processes through which the values of variables are generated (Bollen, 1989; Pearl, 2000). In structural equation modeling, special types of equations, known as structural equations, are used to represent how the values of variables are determined. An illustrative example of a structural equation for the case described above is given by

$$y = f_y(x, e_y), \tag{1}$$

where y denotes whether the disease is cured (1: cured, 0: not cured), x denotes the presence or absence of medication (1: presence, 0: absence), and e_y denotes all the factors other than x that could contribute to determining the value of y, even when x is held constant. Structural equations represent more than simply mathematical equality. In Eq. (1), the left-hand side of the equation is defined by the right-hand side, i.e., the value of y is completely determined by that of x and e_y through the deterministic function f_y.

Similarly, when defining the structural equation relating to x, we obtain a full description of the data-generating process of the variables x and y, i.e., their SEM, as follows:

$$x = e_x \tag{2}$$

$$y = f_y(x, e_y), \tag{3}$$

where e_x denotes all the factors that could contribute to determining the value of x. In these equations, first the value of e_x is somehow generated, and then the value of x is determined from that of e_x by means of the identity function. Subsequently, the value of e_y is somehow generated, and then the value of y is determined from that of x and e_y through the function f_y. The variables e_x and e_y are known as exogenous variables, external influences, disturbances, errors or background variables. The values of these variables are generated outside of the model and their data-generating processes are decided by the modeler not to be further modeled. In contrast, variables whose values are generated inside the model, such as y above, are known as endogenous variables.

In order to clarify the meanings of SEMs, the qualitative relations are often graphically represented by graphs called path-diagrams. Path-diagrams, also known as causal graphs, can be seen as representing causal structures. Causal graphs are constructed according to two rules (Bollen, 1989; Pearl, 2000): i) Draw a directed edge from every variable on the right-hand side of a structural equation to the variable on the left-hand side; and ii) Draw a bi-directed arc between two exogenous variables if the values of these variables could be (partially) determined by a common latent variable; e.g., in the example above, the level of severity of the disease could contribute to determining both whether the medicine is taken and whether the disease is cured. Common latent variables such as these are called latent confounding variables, and cause the exogenous variables to be dependent. The associated causal graph of the SEM represented by Eq. (2)–(3) is shown in the left of Fig. 1. Since x is determined by e_x, and y could be determined by x and e_y, directed edges are drawn from e_x to x, and from x and e_y to y. Since there could be a common latent variable that contributes to determining the values of both x and y, a bi-directed arc is drawn between e_x and e_y.

In general, a SEM is defined as a four-tuple consisting of i) endogenous variables; ii) exogenous variables; iii) deterministic functions that define the structural equations relating the endogenous and exogenous variables; and iv) the probability distribution

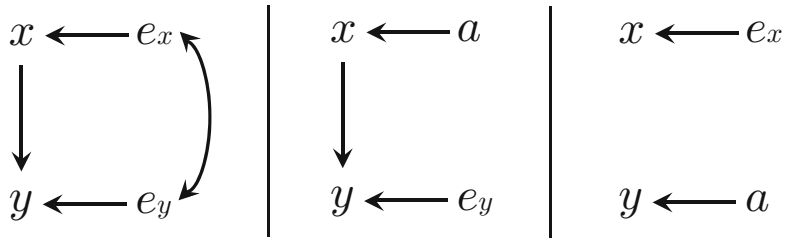

Figure 1: Left: The associated causal graph of the SEM in Eq. (2)–(3). Center: The causal graph after intervening on x in the left-most SEM. Right: The causal graph after intervening on y in the left-most SEM.

of the exogenous variables (Pearl, 2000). The probability distribution of the endogenous variables is induced by the deterministic functions and the probability distribution of the exogenous variables. We are able to make inferences on the SEM based on the distribution of the observed variables among the exogenous and endogenous variables. In the example above, the SEM given in Eq. (2)–(3), with the causal graph shown on the left of Fig. 1, consists of i) the endogenous variable y; ii) the exogenous variables $e_x (= x)$ and e_y; iii) the deterministic function f_y; and iv) the probability distribution of the exogenous variables $p(e_x, e_y)$.

2.3 SEM representation of causation

In this subsection, we explain the SEM representation of population-level causation (Pearl, 2000). We first define interventions in SEMs. Intervening on a variable x means holding the variable x to be a constant, a, regardless of the other variables, and this intervention is denoted by $do(x = a)$. In structural equation modeling, this means replacing the function determining x with the constant a, i.e., letting all the individuals in a population take $x = a$ (Pearl, 2000). Suppose that we intervene on x and fix x at a in the example given in Eq. (2)–(3). We then obtain a new SEM, denoted by $M_{x=a}$:

$$x = a \tag{4}$$

$$y = f_y(x, e_y). \tag{5}$$

As a result, the causal graph changes to that shown in the center of Fig. 1. The exogenous variable x becomes independent of the exogenous variable e_y, i.e., the bi-directed arc in the causal graph of the original SEM given in Eq. (2)–(3) disappears, since x is forced to be a regardless of the other variables. Note that we assume that, even if a function is replaced with a constant, the other functions do not change, although this might be physically unrealistic in some cases. In our example, the revised SEM given in Eq. (4)–(5) represents a hypothetical population, where all the individuals in the population are forced to take $x = a$, but the other function f_y, which relates x to y, does not change.

Next, we define post-intervention distributions (Pearl, 2000). When intervening on

x, the post-intervention distribution of y is defined by the distribution of y in the SEM after the intervention $M_{x=a}$:

$$p(y|do(x=a)) := p_{M_{x=a}}(y). \tag{6}$$

In the example above, the post-intervention distribution of y (1: cured, 0: not cured) when fixing x at a (1: taking the medicine, 0: not taking the medicine) is given by the distribution of y in the post-intervened SEM $M_{x=a}$, for which the associated causal graph is shown in the center of Fig. 1.

We can now provide the SEM representation of population-level causation (Pearl, 2000). If there exist two different values c and d, such that the post-intervention distributions are different; that is,

$$p(y|do(x=c)) \neq p(y|do(x=d)), \tag{7}$$

we can say that x causes y in this population. In the example we are using, if $p(y|do(x=1)) \neq p(y|do(x=0))$, we can say that taking the medicine positively or negatively causes a cure in this population. Moreover, if $p(y=1|do(x=1)) > (<)\ p(y=1|do(x=0))$, we can say that taking the medicine positively (negatively) causes, i.e., is effective (harmful) in curing the disease in this population.

A common method for quantifying the causal connection strength of x on y is to assess the following average difference (Rubin, 1974; Pearl, 2000):

$$E(y|do(x=d)) - E(y|do(x=c)), \tag{8}$$

which is called the average causal effect. This evaluates to what extent, on average, the value of y will change if the value of x is changed from c to d. Changing the value of x from c to d means that x is fixed at c, regardless of the variables that determine x, and the value is changed from c to d (Pearl, 2000). As explained above, fixing x at c, regardless of the variables that determine x, the process that is denoted by $do(x=c)$, means replacing the function determining x with c in the SEM.

Although x and y are binary, purely for the purpose of illustration, we assume that the function f_y, in the SEM of Eq. (2)–(3), is linear:

$$x = e_x \tag{9}$$
$$y = b_{yx}x + e_y, \tag{10}$$

where b_{yx} is constant. The post-intervened SEM $M_{x=a}$ takes the form:

$$x = a \tag{11}$$
$$y = b_{yx}x + e_y. \tag{12}$$

Therefore, the average causal effect of x on y when x is changed from c to d is given by

$$E(y|do(x=d)) - E(y|do(x=c)) = E(b_{yx}d + e_y) - E(b_{yx}c + e_y) \tag{13}$$

$$= b_{yx}(d - c). \tag{14}$$

The expected average change in x is thus the difference between d and c multiplied by the coefficient b_{yx}, while the post-intervened model $M_{y=a}$ shown on the right of Fig. 1 is written as

$$x = e_x \tag{15}$$

$$y = a. \tag{16}$$

Then, the average causal effect of y on x when changing y from c to d is given by

$$E(x|do(y = d)) - E(x|do(y = c)) = E(e_x) - E(e_x) \tag{17}$$

$$= 0. \tag{18}$$

This is reasonable, since y does not contribute to defining x in the original SEM shown in Eq. (2)–(3) and on the left of Fig. 1.

Structural equation models can also be used to represent individual-level causation. The key concept in such a situation is that different values of the vectors that collect exogenous variables can be seen as representing different individuals (Pearl, 2000).

The values of e_x and e_y for Taro in the medicine cure example in Eq. (2)–(3) are denoted by e_x^{Taro} and e_y^{Taro}, respectively. Furthermore, the values that y would attain had x been fixed at d and c are denoted by $y_{x=d}^{Taro}$ and $y_{x=c}^{Taro}$. The values $y_{x=d}^{Taro}$ and $y_{x=c}^{Taro}$ are obtained as the solutions of the SEMs $M_{x=d}$ with x fixed at d and $M_{x=c}$ with x fixed at c when the values of the exogenous variables e_x and e_y are e_x^{Taro} and e_y^{Taro}. The difference between $y_{x=d}^{Taro}$ and $y_{x=c}^{Taro}$ is thus

$$y_{x=d}^{Taro} - y_{x=c}^{Taro} = f_y(d, e_y^{Taro}) - f_y(c, e_y^{Taro}). \tag{19}$$

If there exist two different values, c and d, such that the difference is not zero, we can say that x causes y for Taro. This means that, if x for Taro is changed from c to d, y for Taro increases by $f_y(d, e_y^{Taro}) - f_y(c, e_y^{Taro})$. This can be simplified to $b_{yx}(d - c)$ if f_y is linear, which means that if x for Taro is changed from c to d, y for Taro increases by the difference between d and c multiplied by the coefficient b_{yx}.

2.4 Identifiability of average causal effects when the causal structure is known

So far, we have provided definitions for various causal concepts. We now briefly discuss the identifiability conditions required for average causal effects to be uniquely estimated from the observed data when the causal structure is known. We consider the situation where $E(y|do(x))$ is reduced to an expression without any $do(\cdot)$ operators.

In the simplest case, the relation of x and y is acyclic, i.e., there is no directed cycle in the causal structure, and the exogenous variables are independent, which implies that there are no latent confounders:

$$x = e_x \tag{20}$$

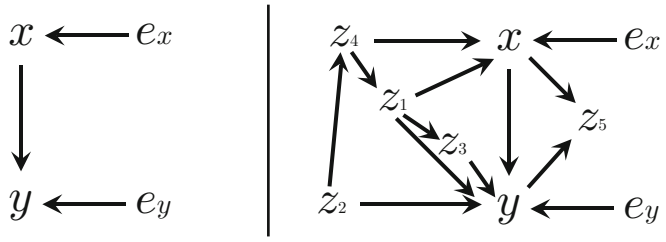

Figure 2: Left: The causal graph of the SEM in Eq. (20)–(21). Right: An example of a causal graph where observing z_1 and z_4 is sufficient for identifying the causal effect from x to y. The exogenous variables corresponding to z_q $(q = 1, \cdots, 5)$ are omitted from the illustration.

$$y = f_y(x, e_y), \tag{21}$$

where exogenous variables e_x and e_y are independent, in contrast to the SEM in Eq. (2)–(3). If some latent confounders do exist, this means the exogenous variables are dependent. The causal structure of the model is shown on the left of Fig. 2. In this case, it can straightforwardly be shown that $E(y|do(x)) = E(y|x)$ (Pearl, 1995). Following this, the average causal effect is calculated by the difference between two conditional expectations:

$$E(y|do(x = d)) - E(y|do(x = c)) = E(y|x = d) - E(y|x = c). \tag{22}$$

We can also describe a more general case, where the additional variables z_q $(q = 1, \cdots, Q)$ exist. Assume that the causal relations of x, y and z_q $(q = 1, \cdots, Q)$ are acyclic, and their exogenous variables are independent. It must now be decided which of the variables z_i should be observed and used to identify $E(y|do(x))$. A sufficient set of variables for this is that of the parents of x, i.e., the variables that have directed edges to x (Pearl, 1995). Then, the average causal effect can be estimated by

$$\begin{aligned} &E(y|do(x = d)) - E(y|do(x = c)) \\ &\quad = E_{pa(x)}[E(y|x = d, pa(x))] - E_{pa(x)}[E(y|x = c, pa(x))], \end{aligned} \tag{23}$$

where $pa(x)$ denotes the set of parents of x. If f_y is linear, the average causal effect can be simplified to the difference $(d - c)$ multiplied by the partial regression coefficient of x when y is regressed on x and its parents. An example of a causal structure is given on the right of Fig. 2, where observing z_1 and z_4 is sufficient. Further details regarding latent confounder cases can be found in Shpitser and Pearl (2006, 2008). Once the causal structure is known, in many cases it is possible to determine whether average causal effects are identifiable, i.e., can be uniquely estimated from the observed data.

2.5 Identifiability of causal structures

In this subsection, we discuss the identifiability of causal structures, i.e., under which model assumptions the causal structure of variables can be uniquely estimated

based on the observed data. Model assumptions represent the background knowledge and hypotheses of the modeler and place constraints on the SEM. These assumptions can sometimes be tested to detect possible violations, although, as in any data analysis process, it would be impossible to prove that they are true.

2.5.1 Basic setup

We first explain the basic setup for identifying causal structures (Pearl, 2000; Spirtes et al., 1993). We assume that the causal relations of the observed variables are acyclic, i.e., there are no directed cycles or feedback loops in the causal graph. Since the exogenous variables are independent, it is implied that there are no latent or unobserved confounding variables that causally influence more than one variable. Although these assumptions may appear to be restrictive, it is possible to relax the two assumptions and develop more general methods based on the information obtained from the basic setup.

In this paper, the focus is on continuous variable cases. Although no specific functional form is assumed for discrete-valued data, in most cases, linearity and Gaussianity are assumed for continuous-valued data (Spirtes et al., 1993; Pearl, 2000). This assumption of linearity would, however, almost certainly be violated when analyzing real-world data. Therefore, in theory, nonlinear approaches are probably more suitable for modeling the causal relations of variables. However, it should be noted that, in practice, linear methods can often provide better results when finding qualititative relations including causal directions is necessary (Pe'er and Hacohen, 2011; Hurley et al., 2012), since nonlinear methods usually require very large sample sizes. In the remainder of the paper, we mainly discuss linear methods, but also refer to their nonlinear extensions. In the following sections, we furthermore show that the assumption of Gaussianity actually limits the applicability of causality estimation methods, and that a significant advantage may be achieved by departing from this assumption.

The basic model for continuous observed variables x_i $(i = 1, \cdots, d)$ is therefore formulated as follows: A causal ordering of the variables x_i is denoted by $k(i)$. With this ordering, the causal relations of the variables x_i can be graphically represented by a directed acyclic graph (DAG)[1)], so that no later variable determines, that is, has a directed path to, any earlier variable in the DAG. Further, we assume that the functional relations of the variables are linear. Without loss of generality, the variables x_i are assumed to have zero mean. We thus obtain a linear acyclic SEM with no latent confounders (Wright, 1921; Bollen, 1989):

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \tag{24}$$

where e_i are continuous latent variables that are exogenous, i.e., are not determined inside the model, and b_{ij} are the connection strengths from x_j to x_i. The exogenous variables e_i have zero mean and non-zero variance, and are independent of each other.

[1)] A directed acyclic graph (DAG) is a graph whose edges are all directed and which has no directed cycles

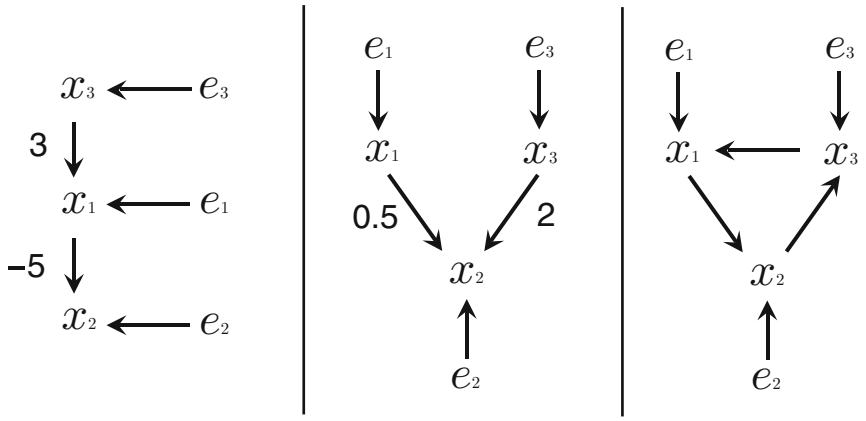

Figure 3: Left and center: Example causal graphs of linear acyclic SEMs. Right: An example causal graph of linear *cyclic* SEMs.

The independence assumption between e_i implies that there are no latent confounding variables.

In matrix form, the linear acyclic SEM with no latent confounders in Eq. (24) can be written as

$$\boldsymbol{x} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}, \tag{25}$$

where the connection strength matrix $\mathbf{B}$ collects the connection strengths b_{ij}, and the vectors $\boldsymbol{x}$ and $\boldsymbol{e}$ collect the observed variables x_i and the exogenous variables e_i, respectively. The zero/non-zero pattern of b_{ij} corresponds to the absence/existence pattern of the directed edges. That is, if $b_{ij} \neq 0$, there is a directed edge from x_j to x_i, but if this is not the case, there is no directed edge from x_j to x_i. Note that, due to the acyclicity, the diagonal elements of $\mathbf{B}$ are all zeros. It can be shown that it is always possible to perform simultaneous, equal row and column permutations on the connection strength matrix $\mathbf{B}$ to cause it to become *strictly* lower triangular, based on the acyclicity assumption (Bollen, 1989). Here, strict lower triangularity is defined as a lower triangular structure with the diagonal consisting entirely of zeros.

Examples of causal graphs for representing the linear acyclic SEMs with no latent confounders in Eq. (25) are provided in Fig. 3. The SEM corresponding to the left-most causal graph of the figure is written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 3 \\ -5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}. \tag{26}$$

In this example, x_3 is in the first position of the causal ordering that causes $\mathbf{B}$ to be strictly lower triangular, x_1 is in the second, and x_2 is in the third, i.e., $k(3) = 1$, $k(1) = 2$, and $k(2) = 3$. If we permute the variables x_1 to x_3 according to the causal ordering, we obtain

$$\begin{bmatrix} x_3 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ 0 & -5 & 0 \end{bmatrix} \begin{bmatrix} x_3 \\ x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_1 \\ e_2 \end{bmatrix}. \tag{27}$$

It can be seen that the resulting connection strength matrix is strictly lower triangular. There is no other such causal ordering of the variables that results in a strictly lower triangular structure in this example. In contrast, there are two such causal orderings in the center causal graph: i) $k(1) = 1$, $k(3) = 2$, and $k(2) = 3$; and ii) $k(3) = 1$, $k(1) = 2$, and $k(2) = 3$, since there is no directed path between x_1 and x_3.

The goal of identifying causal structures under this basic setup is to estimate the unknown, $\mathbf{B}$, by using only the data $\mathbf{X}$, based on the assumption that $\mathbf{X}$ is randomly sampled from a linear acyclic SEM with no latent confounders, as represented by Eq. (25) above. In other words, we aim to determine which model is true among the class of linear acyclic SEMs with no latent confounders, assuming that the class includes the true one.

2.5.2 A conventional approach

In this section, we first discuss the identifiability problems experienced with conventional methods for estimating $\mathbf{B}$ of the linear acyclic SEM with no latent confounders in Eq. (25). We say that $\mathbf{B}$ is identifiable if and only if $\mathbf{B}$ can be uniquely determined or estimated from the data distribution $p(\boldsymbol{x})$. Once $\mathbf{B}$ is identified, we can estimate the causal structure from the zero/non-zero pattern of its elements, b_{ij}. The connection strength matrix $\mathbf{B}$, together with the distribution of the exogenous variables $p(\boldsymbol{e})$, induces the distribution of the observed variables $p(\boldsymbol{x})$. If $p(\boldsymbol{x})$ are different for different $\mathbf{B}$, it follows that $\mathbf{B}$ can be uniquely determined.

The causal Markov condition is a classical principle used for estimating the causal structure of the linear acyclic SEM with no latent confounders in Eq. (25). For any linear acyclic SEM, the causal Markov condition holds[2] (Pearl and Verma, 1991), as follows: Each observed variable x_i is independent of its non-descendants in the DAG conditional on its parents, i.e., $p(\boldsymbol{x}) = \Pi_{i=1}^{d} p(x_i|pa(x_i))$. If Gaussianity of the exogenous variables is furthermore assumed, conditional independence is reduced to partial uncorrelatedness. Thus, conditional independence between observed variables provides a clue as to what the underlying causal structure is.

It is necessary to make an additional assumption, known as faithfulness (Spirtes et al., 1993) or stability (Pearl, 2000), when making use of the causal Markov condition for estimating the causal structure. In this case, the faithfulness assumption means that the conditional independence of x_i is represented by the graph structure only, i.e., by the zero/non-zero status of b_{ij}, and not by the specific values of b_{ij}. Thus, owing to the faithfulness assumption, certain special cases are excluded, so that no conditional independence of x_i holds other than that derived from the causal Markov condition. The following is an example of faithfulness being violated:

[2] The causal Markov condition holds in general cases including in discrete variable cases and nonlinear cases.

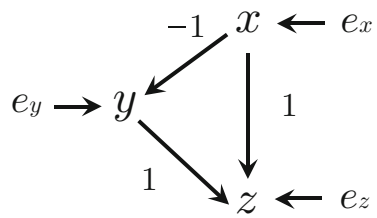

Figure 4: An example of faithfulness being violated.

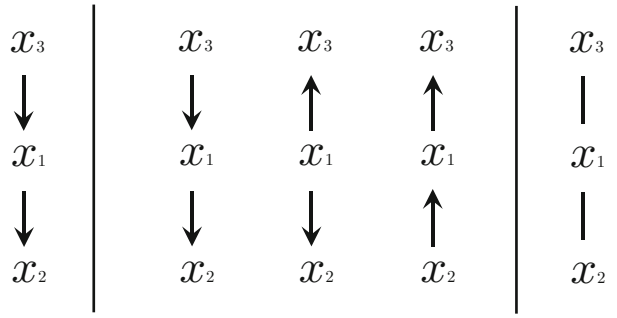

Figure 5: Left: An example of the causal Markov condition being unable to identify the causal structure. Center: The candidate causal structures that give the same conditional independence of variables as the true causal structure on the left. Right: The output based on the causal Markov condition and faithfulness.

$$x = e_x \tag{28}$$

$$y = -x + e_y \tag{29}$$

$$z = x + y + e_z, \tag{30}$$

where e_x, e_y, e_z are Gaussian and mutually independent. The associated causal graph is shown in Fig. 4. When the causal Markov condition is applied to the causal graph, no conditional independence of x_i holds. However, the correlation of x and z is zero, i.e., $\mathrm{cov}(x, z) = 0$, which means that x and z are uncorrelated; in other words, they are independent. Although the faithfulness assumption has often been criticized, it would not be as problematic as in the above case in practice, since such an exact cancellation would rarely occur (Glymour, 2010).

Unfortunately, in many cases, the causal Markov condition used along with faithfulness is not sufficient for uniquely identifying the causal structure of the linear acyclic SEM with no latent confounders in Eq. (25) (Pearl, 2000; Spirtes et al., 1993). An example of this is provided in Fig. 5. Suppose that data $\boldsymbol{x}$ is generated from the causal graph on the left of Fig. 5, but the true causal graph is not known. According to the causal Markov condition, x_2 and x_3 are independent, conditional on x_1, and no other conditional independence holds. The only information available for estimating the underlying causal structure is the conditional independence of x_2 and x_3. Among the class of linear acyclic SEMs with no latent confounders, causal structures that exhibit the same conditional independence as in the data generated from the true

causal graph on the left of Fig. 5 are the three that are shown in the center of Fig. 5. In each of these three causal structures, only x_2 and x_3 are conditionally independent. However, the three causal structures are quite different, and there is no causal direction that is consistent across all the three graphs. The candidate causal structures are usually summarized as shown on the right of Fig. 5, where the undirected edges mean that the directions were not consistent with the candidate graphs. In this example, this is the extent of the estimation that the causal Markov condition and faithfulness are capable of.

Many estimation algorithms based on the causal Markov condition and faithfulness have been proposed (Spirtes et al., 1993; Pearl, 2000). However, many linear acyclic SEMs with no latent confounders exhibit the same set of conditional independence and equally fit the data, as shown in the example above. Even if the Gaussianity of the exogenous variables is assumed in addition (Chickering, 2002), this does not offer a significant advantage. Moreover, many linear acyclic SEMs with no latent confounders show the same Gaussian distribution and equally fit the data, since all of the information is contained in the covariance matrix. For example, consider a comparison of the following two SEMs, with opposing causal directions between the two variables x_1 and x_2:

$$\text{Model 1}: \begin{cases} x_1 = e_1 \\ x_2 = 0.8x_1 + e_2, \end{cases} \tag{31}$$

where e_1 and e_2 are independent, $\mathrm{var}(e_1)$ and $\mathrm{var}(e_2)$ are 1 and 0.6^2 so that $\mathrm{var}(x_1)$ and $\mathrm{var}(x_2)$ are 1s for the sake of simplicity in illustration. Similarly,

$$\text{Model 2}: \begin{cases} x_1 = 0.8x_2 + e_1 \\ x_2 = e_2, \end{cases} \tag{32}$$

where e_1 and e_2 are independent, $\mathrm{var}(e_1)$ and $\mathrm{var}(e_2)$ are 0.6^2 and 1 so that $\mathrm{var}(x_1)$ and $\mathrm{var}(x_2)$ are 1s. In matrix form, the two models may be written as

$$\text{Model 1}: \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} 0 & 0 \\ 0.8 & 0 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}}_{\boldsymbol{e}}, \tag{33}$$

and

$$\text{Model 2}: \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} 0 & 0.8 \\ 0 & 0 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}}_{\boldsymbol{e}}. \tag{34}$$

The connection strength matrices $\mathbf{B}$ of the two models differ to a great extent.

In the above, there are no pairs of variables that are (conditionally) independent, since $\mathrm{cov}(x_1, x_2) = 0.8 \neq 0$ in both of the models. If e_i are furthermore assumed to be Gaussian, the two models provide the same Gaussian distribution of the observed

variables x_1 and x_2 in both models, with the means of the models being zeros, their variables being 1s and their covariance being 0.8. Thus, no distinction can be made between the two models with different causal direction, which means that $\mathbf{B}$ is not identifiable. Similarly, in many cases, the connection strength matrix $\mathbf{B}$ cannot be uniquely identified by using the causal Markov condition and faithfulness.

2.5.3 A non-Gaussian approach

Although the causal Markov condition and Gaussianity assumption were not capable of distinguishing between Models 1 and 2 above, it can be shown that it is possible to distinguish between the two models if the exogenous variables e_1 and e_2 are in fact non-Gaussian and this non-Gaussianity is utilized for model identification (Dodge and Rousson, 2001; Shimizu et al., 2006). We are able to demonstrate that $\mathbf{B}$ in Eq. (25) is identifiable if the independent exogenous variables e_i are non-Gaussian (Shimizu et al., 2006). If the exogenous variables e_1 and e_2 are Gaussian, the distributions of the observed variables do not differ between Models 1 and 2 above, with opposite causal directions existing between x_1 and x_2, as shown in the center of Fig. 6. However, if the exogenous variables e_1 and e_2 are non-Gaussian, and in this case uniformly distributed, the distributions of the observed variables differ between the two models, as shown in the right-most column of the figure. This observation can be generalized to any non-Gaussian distributions of exogenous variables (Shimizu et al., 2006). In the following sections, we explain in more detail the concepts and methods underlying a non-Gaussian approach such as this.

3. LiNGAM

Shimizu et al. (2006) proposed a non-Gaussian version of the linear acyclic SEM with no latent confounders in Eq. (24), known as a linear non-Gaussian acyclic model, abbreviated as LiNGAM:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \tag{35}$$

where e_i are continuous latent variables that are exogenous, and b_{ij} are the connection strengths from x_j to x_i. With the causal ordering of the variables x_i, denoted by $k(i)$, the causal relations of the variables x_i can be graphically represented by using a DAG. The exogenous variables e_i follow *non-Gaussian* distributions, with zero mean and non-zero variance, and are independent of each other. The independence assumption between e_i implies that there are no latent confounding variables. In matrix form, the LiNGAM model in Eq. (35) is written as

$$\boldsymbol{x} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}, \tag{36}$$

where the connection strength matrix $\mathbf{B}$ collects the connection strengths b_{ij}, and the vectors $\boldsymbol{x}$ and $\boldsymbol{e}$ collect the observed variables x_i and the exogenous variables e_i, respectively. Note that the matrix $\mathbf{B}$ can be permuted to become lower triangular

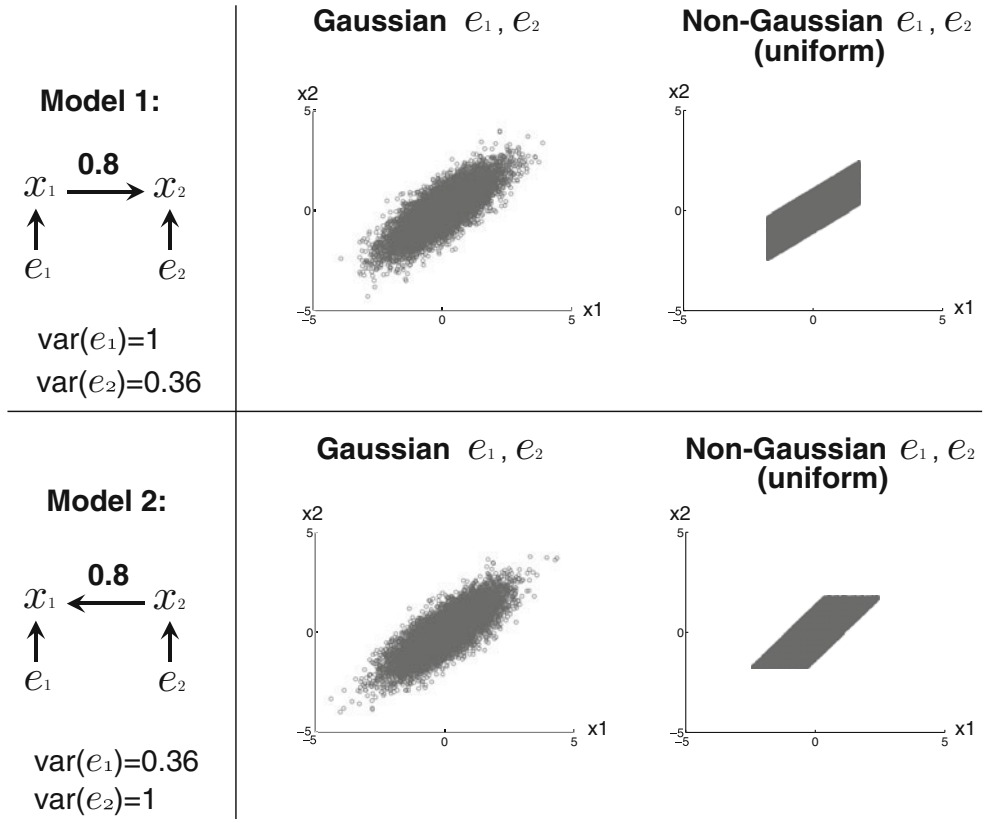

Figure 6: A demonstration of the usefulness of the non-Gaussianity of data.

with all zeros on the diagonal, i.e., strictly lower triangular, if simultaneous, equal row and column permutations are made according to the causal ordering $k(i)$, due to the acyclicity. The difference between this model and the basic model in Eq. (24) is that the exogenous variables e_i are assumed to be *non-Gaussian*. LiNGAM has been proven to be identifiable (Shimizu et al., 2006), i.e., the connection strength matrix $\mathbf{B}$ can be uniquely identified based on the data $\boldsymbol{x}$ only.

3.1 Independent component analysis

Since the concept of independent component analysis (ICA) is closely related to the identifiability of LiNGAM and its estimation, before discussing the identifiability of LiNGAM, we provide a brief overview of ICA (Jutten and Hérault, 1991; Hyvärinen et al., 2001). ICA is a non-Gaussian variant of factor analysis, and the ICA model (Jutten and Hérault, 1991; Comon, 1994) for the observed variables x_i $(i = 1, \cdots d)$ can be defined as follows:

$$x_i = \sum_{j=1}^{d} a_{ij} s_j, \tag{37}$$

where s_j are continuous latent variables that are mutually independent. The latent independent variables s_j are known as independent components of the model, and follow non-Gaussian distributions. The ICA model represents the data-generating process, where the latent independent components s_j are summed with the coefficients a_{ij} and are observed as x_i. In matrix form, the ICA model in Eq. (37) may be represented by

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s}, \tag{38}$$

where the mixing matrix $\mathbf{A}$ collects the coefficients a_{ij}, and the vectors $\boldsymbol{x}$ and $\boldsymbol{s}$ collect the observed variables x_i and the independent components s_j, respectively. The mixing matrix $\mathbf{A}$ is square, i.e., the number of observed variables is equal to the number of independent components, and it is assumed to be of full column rank. It can be shown that, because non-Gaussianity of data is utilized for model identification, $\mathbf{A}$ is identifiable up to the permutation, scaling and sign of the columns, meaning there is no factor rotation indeterminacy (Comon, 1994; Eriksson and Koivunen, 2004). Thus, the mixing matrix identified by ICA $\mathbf{A}_{ICA}$ can be written as

$$\mathbf{A}_{ICA} = \mathbf{APD}, \tag{39}$$

where $\mathbf{P}$ is an unknown permutation matrix and $\mathbf{D}$ is an unknown diagonal matrix with no zeros on the diagonal.

The majority of ICA estimation methods estimate a matrix known as the separating matrix $\mathbf{W} = \mathbf{A}^{-1}$ (Hyvärinen et al., 2001). Furthermore, most of these methods minimize mutual information (or its approximation) of estimated independent components $\hat{\boldsymbol{s}} = \mathbf{W}_{ICA}\boldsymbol{x}$, i.e., $I(\hat{\boldsymbol{s}}) = \{\sum_{j=1}^{d} H(\hat{s}_j)\} - H(\hat{\boldsymbol{s}})$, where $H(\hat{\boldsymbol{s}})$ is the differential entropy of $\hat{\boldsymbol{s}}$ defined by $E\{-\log p(\hat{\boldsymbol{s}})\}$. It can be shown that the mutual information of these estimated independent components is zero if and only if they are independent. Following this, the separating matrix $\mathbf{W}$ is estimated up to the permutation $\mathbf{P}$, and scaling and sign $\mathbf{D}$ of the rows

$$\mathbf{W}_{ICA} = \mathbf{PDW}(= \mathbf{PDA}^{-1}). \tag{40}$$

ICA estimation methods provide a random permutation of the rows. Consistent and computationally efficient estimation algorithms that do not need to specify the distributions of independent components have also been developed (Amari, 1998; Hyvärinen, 1999). Refer to Hyvärinen et al. (2001) and Hyvärinen (2013) for more details on ICA.

3.2 Identifiability of LiNGAM

We now explain the method for identifying the connection strength matrix $\mathbf{B}$ of the LiNGAM in Eq. (36), as provided by Shimizu et al. (2006). Let us first solve Eq. (36) for $\boldsymbol{x}$. From this, we obtain

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{e}, \tag{41}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$. Since the components of $\boldsymbol{e}$ are independent and non-Gaussian,

Eq. (41) defines the ICA model, which, as stated above, is known to be identifiable.

Essentially, ICA is capable of estimating $\mathbf{A}$ (and $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{I} - \mathbf{B}$); however, it exhibits permutation, scaling and sign indeterminacies. ICA gives $\mathbf{W}_{ICA}$=$\mathbf{PDW}$, where $\mathbf{P}$ is an unknown permutation matrix, and $\mathbf{D}$ is an unknown diagonal matrix. However, in LiNGAM, the correct permutation matrix $\mathbf{P}$ can be found (Shimizu et al., 2006): the correct $\mathbf{P}$ is the only one that contains no zeros in the diagonal of $\mathbf{DW}$, since $\mathbf{B}$ should be a matrix that can be permuted to become lower triangular with all zeros on the diagonal and $\mathbf{W} = \mathbf{I} - \mathbf{B}$. Furthermore, the correct scaling and signs of the independent components can be determined by using the unity on the diagonal of $\mathbf{W}$=$\mathbf{I}$−$\mathbf{B}$. To obtain $\mathbf{W}$ it is only necessary to divide the rows of $\mathbf{DW}$ by its corresponding diagonal elements. Finally, the connection strength matrix $\mathbf{B} = \mathbf{I} - \mathbf{W}$ may be computed. It should be noted that we do not assume that the distribution of $\mathbf{x}$ is faithful to the generating graph (Spirtes et al., 1993; Pearl, 2000), unlike in the conventional approach explained in Section 2.5.2.

To illustrate the concept of determining the correct permutation, consider the following LiNGAM model:

$$x_1 = e_1 \tag{42}$$
$$x_2 = b_{21}x_1 + e_2 \tag{43}$$
$$x_3 = b_{32}x_2 + e_3, \tag{44}$$

where e_1, e_2 and e_3 are non-Gaussian and mutually independent. In matrix form, the example model above can be written as follows:

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}_{\boldsymbol{e}}. \tag{45}$$

Rewriting this in the form of ICA, we obtain

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix} \right)^{-1}}_{(\mathbf{I}-\mathbf{B})^{-1}} \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}_{\boldsymbol{e}} \tag{46}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -b_{21} & 1 & 0 \\ 0 & -b_{32} & 1 \end{bmatrix}^{-1}}_{\mathbf{W}^{-1}} \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}_{\boldsymbol{e}}. \tag{47}$$

In this case, the correct $\mathbf{W} = \mathbf{I} - \mathbf{B}$ is

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 \\ -b_{21} & 1 & 0 \\ 0 & -b_{32} & 1 \end{bmatrix}, \tag{48}$$

which is lower triangular and contains no zeros in the diagonal. Premultiplying $\mathbf{W}$ by a diagonal matrix $\mathbf{D}$ with no zeros in the diagonal does not have an effect on the zero/non-zero pattern of $\mathbf{W}$, since

$$\mathbf{DW} = \begin{bmatrix} d_{11} & 0 & 0 \\ -d_{22}b_{21} & d_{22} & 0 \\ 0 & -d_{33}b_{32} & d_{33} \end{bmatrix}. \tag{49}$$

However, any other permutation of the rows of $\mathbf{DW}$ does affect the zero/non-zero pattern of $\mathbf{DW}$, and introduces a zero into the diagonal. To demonstrate this, we show that, by exchanging the first and second rows, which is represented by the permutation matrix $\mathbf{P}^{12}$, we obtain

$$\mathbf{P}^{12}\mathbf{DW} = \begin{bmatrix} -d_{22}b_{21} & d_{22} & 0 \\ d_{11} & 0 & 0 \\ 0 & -d_{33}b_{32} & d_{33} \end{bmatrix}, \tag{50}$$

which contains a zero in the diagonal. Therefore, by making use of this approach, we can determine the correct permutation matrix $\mathbf{P}$ by finding a permutation matrix such that the permuted matrix contains no zeros in the diagonal.

We can thus conclude that no condition on e_i other than non-Gaussianity is required for LiNGAM to be identifiable (Shimizu et al., 2006), similarly to ICA (Comon, 1994; Eriksson and Koivunen, 2004). However, for the estimation methods to be consistent, additional assumptions, e.g., the existence of their moments or some other statistics, must be made in order to ensure that the statistics computed in the estimation algorithms do in fact exist.

4. Estimation of LiNGAM

The log likelihood of LiNGAM in Eq. (36) for a given causal ordering $k(i)$ (Hyvärinen et al., 2010) is represented by

$$\log L(\mathbf{X}) = \sum_t \sum_i \log p_i \left(\frac{\boldsymbol{x}(t) - \boldsymbol{b}_i^T \boldsymbol{x}(t)}{\sigma_i} \right) - n \sum_i \log \sigma_i, \tag{51}$$

where $\mathbf{X}$ is the observed sample, $\boldsymbol{x}(t)$ are the t-th observations, $\boldsymbol{b}_i^T$ are the i-th row vectors of $\mathbf{B}$, σ_i are the standard deviations of e_i, n is the number of observations and $p_i = p(e_i/\sigma_i)$ are the probability densities of the standardized versions of e_i, i.e., e_i/σ_i.

A straightforward approach would be to estimate the connection strength matrix $\mathbf{B}$, which maximizes the likelihood over all the possible causal orderings $k(i)$. However, such an approach would not be adequate (Hyvärinen et al., 2010), as it would be extremely costly computationally, since the number of possible causal orderings increases very quickly when large numbers of variables are involved. In principle, we could estimate the densities p_i, but it is preferable to avoid this approach if possible.

Thus, two estimation algorithms (Shimizu et al., 2006, 2011) have been proposed, in which it is not necessary to investigate all the possible causal orderings or estimate their probability densities. Both of the approaches estimate a causal ordering of variables $k(i)$ that causes the connection strength matrix $\mathbf{B}$ to become strictly lower triangular. The existence of such a causal ordering of variables is ensured by the assumption of acyclicity (Bollen, 1989). Once a causal ordering of variables is found in this way, we can prune redundant connection strengths b_{ij} (directed edges); that is, find actual zero coefficients by using ordinary sparse regression methods, including that of the adaptive lasso (Zou, 2006)[3]. Zhang and Chan (2006) proposed combining the two steps of finding a causal ordering and pruning redundant connection strengths into one by applying ICA with sparse coefficients.

4.1 ICA-LiNGAM

The first estimation algorithm for LiNGAM, ICA-LiNGAM (Shimizu et al., 2006), involves the same process of demonstrating identifiably, i.e., first, ICA is applied, and second, the estimated separating matrix is permuted so that the diagonal elements of the permuted separating matrix are as large in absolute value as possible; and finally, a causal ordering of variables is found that makes the permuted separating matrix as close to being strictly lower triangular as possible. The ICA-LiNGAM algorithm provided by Shimizu et al. (2006) is described as follows:

ICA-LiNGAM algorithm:

1. Given a d-dimensional random vector $\mathbf{x}$ and its $d \times n$ observed data matrix $\mathbf{X}$, apply an ICA algorithm to obtain an estimate of $\mathbf{A}$.
2. Find the unique permutation of the rows of $\mathbf{W}{=}\mathbf{A}^{-1}$ that yields a matrix $\widetilde{\mathbf{W}}$ without any zeros on the main diagonal. The permutation is sought by minimizing $\sum_i 1/|\widetilde{\mathbf{W}}_{ii}|$. This minimization problem is the classical linear assignment problem, and here the Hungarian algorithm (Kuhn, 1955) is used.
3. Divide each row of $\widetilde{\mathbf{W}}$ by its corresponding diagonal element in order to yield a new matrix $\widetilde{\mathbf{W}}'$ with a diagonal consisting entirely of 1s.
4. Compute an estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ by using $\widehat{\mathbf{B}} = \mathbf{I} - \widetilde{\mathbf{W}}'$.
5. Finally, to estimate a causal order $k(i)$, determine the permutation matrix $\widetilde{\mathbf{P}}$ of $\widehat{\mathbf{B}}$, obtaining the matrix $\widetilde{\mathbf{B}} = \widetilde{\mathbf{P}}\widehat{\mathbf{B}}\widetilde{\mathbf{P}}^T$ that is as close as possible to having a strictly lower triangular structure. For a small number of variables, i.e., fewer than 8, the lower triangularity of $\widetilde{\mathbf{B}}$ can be measured by using the sum of squared b_{ij} in its upper triangular section $\sum_{i \le j} \widetilde{b}_{ij}^2$. In addition, an exhaustive search over all possible permutations is feasible and is hence performed. For higher-dimensional data, the following approximate algorithm is used, which sets small absolute valued elements in $\widetilde{\mathbf{B}}$ to zero, and whereby it can be determined whether it is

---

[3] Redundant connection strengths (directed edges) b_{ij} can be pruned by repeatedly applying adaptive lasso (Zou, 2006) on each variable and its potential parents, for example (Shimizu et al., 2011).

possible to permute the resulting matrix to become strictly lower triangular:

(a) Set the $d(d+1)/2$ smallest (in absolute value) elements of $\widehat{\mathbf{B}}$ to zero.
(b) Repeat
   i. Determine whether $\widehat{\mathbf{B}}$ can be permuted to become strictly lower triangular. If this is possible, stop and return the permuted $\widehat{\mathbf{B}}$; that is, $\widetilde{\mathbf{B}}$.
   ii. In addition, set the next smallest (in absolute value) element of $\widehat{\mathbf{B}}$ to zero.

The ICA-LiNGAM algorithm is computationally efficient, owing to the availability of well-developed ICA techniques. However, this algorithm has a potential downfall, in that most ICA algorithms, including FastICA (Hyvärinen, 1999) and gradient-based algorithms (Amari, 1998), may converge to local optima if the initially guessed state is not properly chosen (Himberg et al., 2004), or if the step size is not suitably selected in gradient-based methods. The appropriate selection of such algorithmic parameters is therefore a complex task.

4.2 DirectLiNGAM

The second estimation algorithm for LiNGAM is known as DirectLiNGAM (Shimizu et al., 2011). DirectLiNGAM is an alternative estimation method that does not make use of ICA. In contrast to ICA-LiNGAM, the DirectLiNGAM algorithm is guaranteed to converge to the right solution in a fixed number of steps, which are equal to the number of variables, provided that all of the model assumptions are met and the sample size is infinite. DirectLiNGAM estimates a causal ordering of variables $k(i)$ that results in the connection strength matrix $\mathbf{B}$ to becoming strictly lower triangular. Once such a causal ordering of variables is found, it is possible to determine actual zero connection strengths by using ordinary sparse regression methods (Zou, 2006), similarly to the process followed in ICA-LiNGAM.

To illustrate the concept underlying DirectLiNGAM, we consider the following example:

$$\begin{bmatrix} x_3 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1.5 & 0 & 0 \\ 0 & -1.3 & 0 \end{bmatrix} \begin{bmatrix} x_3 \\ x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_1 \\ e_2 \end{bmatrix}, \tag{52}$$

where e_1, e_2 and e_3 are non-Gaussian and independent. The procedure of DirectLiNGAM is illustrated in Fig. 7. In DirectLiNGAM, first an exogenous variable is found, which is a variable that is not determined inside the model, i.e., has no parents in the model (Bollen, 1989), and the corresponding row of $\mathbf{B}$ contains only zeros. In the example given in Eq. (52) above, x_3 is an exogenous variable and the corresponding row of $\mathbf{B}$, i.e., the first row, consists entirely of zeros. Therefore, the exogenous variable $x_3 (= e_3)$ can be at the top of a causal ordering such as this that

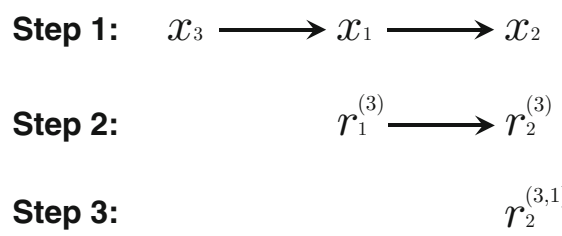

Figure 7: An illustration of DirectLiNGAM: $r_2^{(3,1)}$ denotes the residual when $r_2^{(3)}$ is regressed on $r_1^{(3)}$.

causes $\mathbf{B}$ to be lower triangular with zeros on the diagonal. Following this, the effect of the exogenous variable x_3 is removed from the other variables by using least-squares regression. In other words, we compute the residuals $r_i^{(3)}$ when the other variables x_i $(i = 1, 2)$ are regressed on the exogenous x_3. It can be shown that the residuals $r_i^{(3)}$ $(i = 1, 2)$ follow a LiNGAM model if the relevant assumptions are met and the sample size is infinite (Shimizu et al., 2011). Thus, we have

$$\begin{bmatrix} r_1^{(3)} \\ r_2^{(3)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1.3 & 0 \end{bmatrix} \begin{bmatrix} r_1^{(3)} \\ r_2^{(3)} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \tag{53}$$

The causal ordering of the residuals $r_1^{(3)}$ and $r_2^{(3)}$ is equivalent to that of the corresponding original variables x_1 and x_2. Following this, DirectLiNGAM determines an exogenous residual, in this case, $r_1^{(3)}$. This implies that its corresponding original variable x_1 can be in the second position of the causal ordering, and the remaining variable, x_2, will then be third in the causal ordering. According to this method, DirectLiNGAM estimates the causal orders of variables one by one, from the top downwards.

We now describe a principle that can be used to identify an exogenous variable. We quote the Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953), since this is used to prove Lemma 1 below, which is detailed following this.

Theorem 1 (Darmois-Skitovitch theorem) *Define two random variables, y_1 and y_2, as linear combinations of independent random variables s_i(i=1, $\cdots$, Q):*

$$y_1 = \sum_{i=1}^{Q} \alpha_i s_i, \quad y_2 = \sum_{i=1}^{Q} \beta_i s_i.$$

Then, it can be shown that, if y_1 and y_2 are independent, all variables s_j for which $\alpha_j \beta_j \neq 0$ are Gaussian. □

This theorem therefore shows that, if there exists a non-Gaussian s_j for which $\alpha_j \beta_j \neq 0$, y_1 and y_2 are dependent.

Lemma 1 (Lemma 1 of Shimizu et al. (2011)) *Assume that all the model assumptions of LiNGAM in Eq. (36) are met and the sample size is infinite. Denote*

by $r_i^{(j)}$ *the residual when* x_i *is regressed on* x_j*:* $r_i^{(j)} = x_i - \frac{\mathrm{cov}(x_i,x_j)}{\mathrm{var}(x_j)}x_j$ $(i \neq j)$*. Then a variable* x_j *is exogenous if and only if* x_j *is independent of its residuals* $r_i^{(j)}$ *for all* $i \neq j$*.* □

To illustrate the meaning of the lemma, we describe the following two variable cases. Firstly, the case where x_1 is exogenous is considered:

$$x_1 = e_1 \tag{54}$$
$$x_2 = b_{21}x_1 + e_2, \tag{55}$$

where $b_{21} \neq 0$. Regressing x_2 on x_1,

$$r_2^{(1)} = x_2 - \frac{\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}x_1 \tag{56}$$
$$= x_2 - b_{21}x_1 \tag{57}$$
$$= e_2. \tag{58}$$

Thus, if $x_1 (= e_1)$ is exogenous, since e_1 and e_2 are independent, x_1 and $r_2^{(1)} (= e_2)$ are also independent.

Next, we consider the case where x_1 is not exogenous:

$$x_1 = b_{12}x_2 + e_1 \tag{59}$$
$$x_2 = e_2, \tag{60}$$

where $b_{12} \neq 0$. Regressing x_2 on x_1,

$$r_2^{(1)} = x_2 - \frac{\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}x_1 \tag{61}$$
$$= x_2 - \frac{\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}(b_{12}x_2 + e_1) \tag{62}$$
$$= \left\{1 - \frac{b_{12}\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}\right\}x_2 - \frac{\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}e_1 \tag{63}$$
$$= \left\{1 - \frac{b_{12}\mathrm{cov}(x_2,x_1)}{\mathrm{var}(x_1)}\right\}e_2 - \frac{b_{12}\mathrm{var}(x_2)}{\mathrm{var}(x_1)}e_1. \tag{64}$$

Thus, if x_1 is not exogenous, according to the Darmois-Skitovitch theorem, x_1 and $r_2^{(1)}$ are dependent, since e_1 and e_2 are non-Gaussian and independent. Furthermore, the coefficient of e_1 on x_1 and that of e_1 on $r_2^{(1)}$ are non-zero, since $b_{12} \neq 0$ by definition. Therefore, exogenous variables can be determined by examining the independence between variables and their residuals.

In practice, an exogenous variable may be identified by determining the variable that is the most independent of its residuals. To evaluate independence, a measure needs to be used that is not restricted to uncorrelatedness, since the result of least-squares regression is residuals that are always uncorrelated with, but not necessarily independent of, explanatory variables. For the same reason, non-Gaussianity

is required for the estimation, as uncorrelatedness is equivalent to independence for Gaussian variables.

A simple approach for evaluating independence is to firstly evaluate the pairwise independence between a variable and each of the residuals, and then take the sum of the pairwise measures over the residuals. The mutual independence of random variables is equivalent to their pairwise independence in linear models with non-Gaussian independent latent variables (Comon, 1994). We use U to denote the set of variable indices of $\boldsymbol{x}$; that is, $U=\{1, \cdots, d\}$. From this, we make use of the following statistic to evaluate the independence between a variable x_j and its residuals $r_i^{(j)} = x_i - \frac{\mathrm{cov}(x_i,x_j)}{\mathrm{var}(x_j)} x_j$ when x_i is regressed on x_j $(j \neq i)$:

$$T(x_j; U) = \sum_{i \in U, i \neq j} I_M(x_j, r_i^{(j)}), \tag{65}$$

where $I_M(x_j, r_i^{(j)})$ is the measure of independence between x_j and $r_i^{(j)}$. It is common to use the mutual information between two variables y_1 and y_2 as a measure of independence between them (Hyvärinen et al., 2001). Many non-parametric independence measures (Bach and Jordan, 2002; Gretton et al., 2005; Kraskov et al., 2004), as well as measures that are computationally more simple, which use a single nonlinear correlation of the form $\mathrm{corr}(g(y_1), y_2)$ ($g(\cdot)$ is a nonlinear function) (Hyvärinen, 1998), have also been proposed. Any such method of independence could potentially be used as $I_M(x_j, r_i^{(j)})$ in Eq. (65).

We now present the DirectLiNGAM algorithm (Shimizu et al., 2011) for estimating a causal ordering in the LiNGAM given in Eq. (36), which repeatedly performs least-squares simple linear regression and the evaluation of pairwise independence between each variable and its residuals:

DirectLiNGAM algorithm:

1. Given a d-dimensional random vector $\boldsymbol{x}$, a set of its variable indices U and a $d \times n$ data matrix of the random vector as $\mathbf{X}$, initialize an ordered list of variables $K := \emptyset$.
2. Repeat until $d-1$ variable indices are appended to K:
   (a) Perform least-squares regressions of x_i on x_j for all $i \in U \backslash K$ $(i \neq j)$ and compute the residual vectors $\boldsymbol{r}^{(j)}$ and the residual data matrix $\mathbf{R}^{(j)}$ from the data matrix $\mathbf{X}$, for all $j \in U \backslash K$. Find a variable x_m that is the most independent of its residuals:

$$x_m = \arg \min_{j \in U \backslash K} T(x_j; U \backslash K),$$

   where T is the independence measure defined in Eq. (65).
   (b) Append m to the end of K.
   (c) Let $\boldsymbol{x} := \boldsymbol{r}^{(m)}$, $\mathbf{X} := \mathbf{R}^{(m)}$.

3. Append the remaining variable index to the end of K.

Note that if the i-th element of K is j, it can be seen that $k(j) = i$.

4.3 Improvements on the basic estimation methods

Several improvements on the basic estimation methods have been proposed. Hyvärinen and Smith (2013) proposed a likelihood-ratio-based method for determining an exogenous variable in the DirectLiNGAM framework, a method which is simpler computationally than DirectLiNGAM, since it only needs to evaluate the one-dimensional differential entropies of variables and residuals, and does not need to evaluate their pairwise independence.

Another direction taken is that of using a divide-and-conquer approach. Cai et al. (2013) proposed a principle of dividing observed variables into smaller subsets, in which variables follow a LiNGAM model under the assumption that the causal structure of all of the variables is sparse. By using this approach, LiNGAM estimation methods can be applied to smaller sets of variables, which leads to more accurate estimations and allows large numbers of variables to be handled more easily.

In Tashiro et al. (2014), DirectLiNGAM was extended in order to be robust against latent confounders. Here, the key concept is to detect latent confounders by testing the independence between estimated exogenous variables, and finding subsets that include variables that are not affected by latent confounders, in order to estimate causal orders one by one, as in DirectLiNGAM.

Hoyer and Hyttinen (2009) and Henao and Winther (2011) proposed Bayesian approaches for learning the basic LiNGAM given in Eq. (36).

4.4 Relation to the causal Markov condition

The following three estimation principles have been shown to be equivalent in terms of the estimation of linear acyclic SEMs with no latent confounders (Zhang and Hyvärinen, 2009a; Hyvärinen et al., 2010): i) Maximization of independence between exogenous variables; ii) Minimization of the sum of entropies of exogenous variables; and iii) the causal Markov condition that each variable is independent of its non-descendants in the DAG conditional on its parent, as well as maximization of independence between the parents of each variable and its corresponding exogenous variables. It is therefore clear that non-Gaussianity is more useful than the causal Markov condition for the estimation process. If exogenous variables are Gaussian, least-squares regression always results in the parents of each variable and its corresponding exogenous variables being independent.

4.5 Evaluation of statistical reliability

In many applications, it is often necessary to assess the statistical reliability or

statistical significance of specific LiNGAM estimation results. Several methods for evaluating reliability, based on bootstrapping (Efron and Tibshirani, 1993), have been proposed (Hyvärinen et al., 2010; Komatsu et al., 2010; Thamvitayakul et al., 2012). If either the sample size or the magnitude of non-Gaussianity is small, LiNGAM analysis would provide significantly different results for different bootstrap samples. Smaller non-Gaussianity causes the model to become closer to not being identifiable. Hyvärinen and Smith (2013) proposed a permutation test to find statistically significant causal connection strengths b_{ij}, using multiple data sets that are measured under different conditions.

4.6 Detection of violations of model assumptions

It is possible to detect violations of the model assumptions that may occur. For example, non-Gaussianity of exogenous variables can be tested by means of Gaussianity tests for estimated exogenous variables, such as the Kolmogorov-Smirnov test. In addition, violations of the independence of exogenous variables may be detected by using the independence test of residuals (Entner and Hoyer, 2011; Tashiro et al., 2014). The overall suitability of the model assumptions can be evaluated by means of a chi-square test, using higher-order moments (Shimizu and Kano, 2008), although large sample sizes are required in order to estimate higher-order moments accurately.

5. Extensions of LiNGAM

In this section, we provide a brief overview of some of the extensions of LiNGAM.

5.1 Latent confounding variables

We first discuss an extension of LiNGAM that applies to cases with latent confounders (unobserved common causes). The authors are of the opinion that this is one of the most important areas that LiNGAM can be extended into.

The independence assumption between e_i in LiNGAM given in Eq. (36) implies that there are no latent confounding variables (Shimizu et al., 2006). A latent confounding variable is an unobserved variable that contributes to determining the values of more than one observed variable (Hoyer et al., 2008b). Latent confounding variables exist in many applications, and if such latent confounders are completely ignored, the estimation results obtained may be seriously biased (Bollen, 1989; Spirtes et al., 1993; Pearl, 2000). For this reason, Hoyer et al. (2008b) proposed LiNGAM with latent confounders, and the model provided can be formulated as follows:

$$\boldsymbol{x} = \mathbf{B}\boldsymbol{x} + \mathbf{\Lambda}\boldsymbol{f} + \boldsymbol{e}, \tag{66}$$

where the difference obtained from LiNGAM in Eq. (36) represents the existence of the latent confounding variable vector $\boldsymbol{f}$. The vector $\boldsymbol{f}$ collects the non-Gaussian latent confounders f_q with zero mean and unit variance ($q = 1, \cdots, Q$). Without

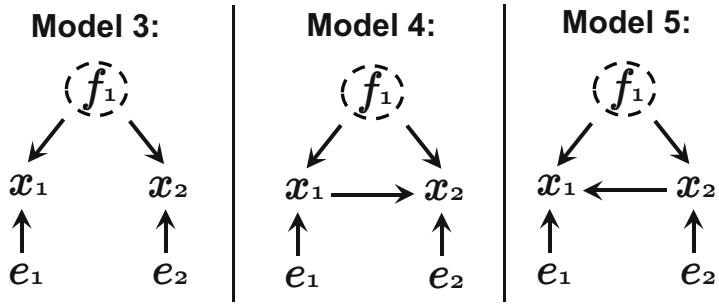

Figure 8: The utilization of non-Gaussianity enables us to distinguish between the three models containing latent confounders. Only one latent confounder is shown in the causal graphs, for the sake of illustration simplicity.

loss of generality, the latent confounders f_q are assumed to be independent of each other, since any dependent latent confounders can be remodeled by means of linear combinations of independent exogenous variables, provided that the underlying model is linear acyclic and the exogenous variables corresponding to the observed variables and latent confounders are independent (Hoyer et al., 2008b). The matrix $\mathbf{\Lambda}$ collects λ_{iq}, which denote the connection strengths from f_q to x_i. It has been shown (Hoyer et al., 2008b) that one can distinguish between the following three models, i.e., the following three different causal structures of observed variables induce different data distributions, when assuming faithfulness of x_i and f_q, and non-Gaussianity of f_q and e_i:

$$\text{Model 3}: \begin{cases} x_1 = \sum_{q=1}^{Q} \lambda_{1q} f_q + e_1 \\ x_2 = \sum_{q=1}^{Q} \lambda_{2q} f_q + e_2, \end{cases} \tag{67}$$

$$\text{Model 4}: \begin{cases} x_1 = \sum_{q=1}^{Q} \lambda_{1q} f_q + e_1 \\ x_2 = b_{21} x_1 + \sum_{q=1}^{Q} \lambda_{2q} f_q + e_2, \end{cases} \tag{68}$$

$$\text{Model 5}: \begin{cases} x_1 = b_{12} x_2 + \sum_{q=1}^{Q} \lambda_{1q} f_q + e_1 \\ x_2 = \sum_{q=1}^{Q} \lambda_{2q} f_q + e_2, \end{cases} . \tag{69}$$

The corresponding causal graphs are provided in Fig. 8.

Hoyer et al. (2008b) furthermore proposed an estimation method based on overcomplete ICA (Lewicki and Sejnowski, 2000); that is, ICA with more latent variables (independent components) than observed variables. However, at present, the overcomplete ICA estimation algorithms that have been developed often become stuck in local optima, and the estimates are not sufficiently reliable (Entner and Hoyer, 2011). Chen and Chan (2013) proposed a simpler approach for estimating LiNGAM with latent confounders, although this method requires the latent confounders f_q to be Gaussian. Henao and Winther (2011) presented a Bayesian approach for estimating LiNGAM with latent confounders, as given in Eq. (66). In addition, Shimizu and Bollen (2013) proposed an alternative Bayesian estimation approach, based on a variant of LiNGAM that incorporates individual-specific effects.

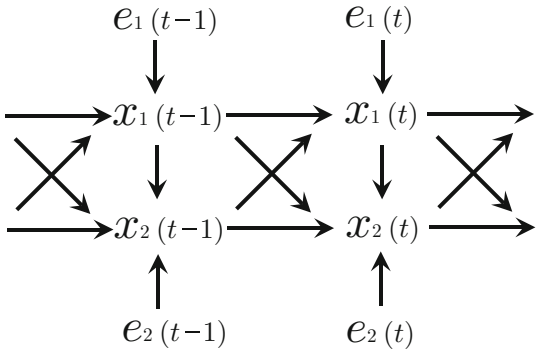

Figure 9: An example of a causal graph of non-Gaussian structural vector autoregressive models.

5.2 Time series

Hyvärinen et al. (2010) considered analyzing both lagged and instantaneous causal effects in time series data, an approach which is both necessary and useful if it is possible that the measurements have a lower time resolution than the causal influences. LiNGAM is used for modeling instantaneous causal effects, while a classic auto-regressive model is used for modeling lagged causal effects, the combination of which leads to the following model:

$$\boldsymbol{x}(t) = \sum_{\tau=0}^{h} \mathbf{B}_\tau \boldsymbol{x}(t-\tau) + \boldsymbol{e}(t), \tag{70}$$

where $\boldsymbol{x}(t)$ and $\boldsymbol{e}(t)$ are the observed variable vectors and the exogenous variable vectors at time point t, respectively. $\mathbf{B}_\tau$ denotes the connection strength matrices having a time lag τ. Note that the time lag τ starts from zero, and $\mathbf{B}_0$ can be permuted to become strictly lower triangular, i.e., the instantaneous causal relations are acyclic. An example causal graph is provided in Fig. 9. The model described above is widely known in econometrics as a structural vector autoregressive model (Swanson and Granger, 1997); however, strong background knowledge of the causal structure is required to identify the model, due to the Gaussianity assumption. Hyvärinen et al. (2010) showed that the model in Eq. (70) is identifiable if $e_i(t)$ are non-Gaussian as well as mutually and temporally independent. A simple estimation method for this model is to fit a classic auto-regressive model on $\boldsymbol{x}(t)$ and apply basic LiNGAM on the residuals (Hyvärinen et al., 2010). Following this, the framework may be further generalized so that it allows lagged and instantaneous latent confounders (Kawahara et al., 2011; Gao and Yang, 2012).

5.3 Cyclic models

Lacerda et al. (2008) and Hyvärinen and Smith (2013) extended LiNGAM to apply to cyclic cases. In such a case, the connection strength matrix $\mathbf{B}$ cannot be permuted

to be lower triangular. Lacerda et al. (2008) provided sufficient conditions for the cyclic model to be identifiable: i) the variables are in equilibrium, i.e., the largest eigenvalue of $\mathbf{B}$ is smaller than 1 in absolute value; ii) the cycles are disjoint; and iii) there are no self-loops. Furthermore, a modified ICA-LiNGAM was proposed as an estimation method for cyclic cases (Lacerda et al., 2008).

5.4 Three-way data models

In some application domains, data are obtained under differing conditions: under different experimental conditions, for different subjects or at different time points. In other words, multiple data sets, or three-way data, are obtained, as opposed to a single data set. Ramsey et al. (2011), Shimizu (2012) and Schaechtle et al. (2013) proposed methods for estimating a common causal ordering or causal structure for multiple data sets. Ramsey et al. (2011) obtained excellent estimation results on simulated functional magnetic resonance imaging (fMRI) data created by Smith et al. (2011). Furthermore, Kadowaki et al. (2013) proposed an approach for estimating time-varying causal structures, based on longitudinal data, which is a type of three-way data where variables are repeatedly measured for the same subjects and at different time points.

5.5 Analysis of groups of variables

Kawahara et al. (2010) proposed a LiNGAM analysis of groups of variables, instead of simply single variables. The authors presented an estimation algorithm for a causal ordering of the groups of variables, so that the groups follow a LiNGAM model. Entner and Hoyer (2012) investigated the possibility of applying such causal analysis of groups of variables to brain-imaging data analysis, where certain background knowledge could be used to divide variables into groups a priori.

5.6 Nonlinear extensions

The concept of LiNGAM has been extended to nonlinear cases (Hoyer et al., 2009; Zhang and Hyvärinen, 2009b; Tillman et al., 2010). Zhang and Hyvärinen (2009b) described the following nonlinear extension of LiNGAM, under the assumptions that the relations were acyclic and there were no latent confounders:

$$x_i = f_{i,2}^{-1}(f_{i,1}(\mathrm{pa}(x_i)) + e_i), \tag{71}$$

where the exogenous variables e_i are independent. Note that $pa(x_i)$ denotes the set of parents of x_i. The authors showed that this model is identifiable with the exception of only a few combinations of functional forms and distributions of exogenous variables (Zhang and Hyvärinen, 2009b; Peters et al., 2011b). These identifiability proofs can be applied to a nonlinear additive SEM with Gaussian exogenous variables, as considered by Imoto et al. (2002). There are ongoing developments in computationally efficient estimation methods for nonlinear models (Mooij et al., 2009; Tillman et al.,

2010; Zhang and Hyvärinen, 2009a,b). Extending these nonlinear models to cover latent confounder cases (Zhang et al., 2010), time series cases (Peters et al., 2013), cyclic cases (Mooij et al., 2011), and discrete variable cases (Peters et al., 2011a) has been investigated.

Before the advent of LiNGAM, the following nonlinear non-parametric version of the linear acyclic SEM with no latent confounders in Eq. (25) was extensively studied (Pearl, 2000; Spirtes et al., 1993):

$$x_i = f_i(\mathrm{pa}(x_i), e_i), \tag{72}$$

where the relations are acyclic and there are no latent confounders. The functional forms of the structural equations remain unspecified. Most of these methods (Pearl and Verma, 1991; Spirtes and Glymour, 1991) make use of the causal Markov condition and faithfulness for model identification. Extensions have also been proposed to cover latent confounder cases (Spirtes et al., 1995), time series cases (Entner and Hoyer, 2010), and cyclic cases (Richardson, 1996). In many cases, these nonlinear non-parametric methods are not capable of uniquely identifying the underlying causal structure; however, it is not necessary for them to make such assumptions as linearity on the functional form.

5.7 Other issues

Shimizu et al. (2009) and Hirayama and Hyvärinen (2011) investigated the causal analysis of latent variables or latent factors, as opposed to observed variables. Hoyer et al. (2008a) proposed a method that is robust against the Gaussianity of exogenous variables. Tillman and Spirtes (2011) and Schölkopf et al. (2012) studied the question of when causal information could be useful for the prediction of associations. Bühlmann et al. (2013) proposed an estimation algorithm for a nonlinear additive SEM with Gaussian exogenous variables (Imoto et al., 2002) and developed its asymptotic theory in a high-dimensional scenario. To the best of our knowledge, no work on selection bias (Spirtes et al., 1995) has yet been undertaken in the context of LiNGAM.

6. Conclusion

Utilization of non-Gaussianity in structural equation modeling is useful for model identification. In this way, a wider variety of causal structures can be estimated than when using classical methods. Non-Gaussian data is encountered in many applications, including the social sciences and the life sciences. The non-Gaussian approach discussed in this paper may be a suitable approach in such applications. Download links to papers and codes on this topic are available on the web[4)].

[4)] `http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/lingampapers.html`

Acknowledgements

S.S was supported by KAKENHI #24700275. We thank tutorial participants at the 40th Annual Meeting of the Behaviormetric Society of Japan (BSJ2012) for interesting discussion, and the chief editor Maomi Ueno for giving us the opportunity to present the tutorial and write this survey. We thank Aapo Hyvärinen, Patrik O. Hoyer, Kento Kadowaki, Naoki Tanaka, the guest editor Jun-ichiro Hirayama and two reviewers for their helpful comments.

References

Amari, S. (1998). Natural gradient learning works efficiently in learning. *Neural Computation*, 10:251–276.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika*, 48:493–517.

Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Bühlmann, P. (2013). Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research*, 77(3):357–370.

Bühlmann, P., Peters, J., and Ernest, J. (2013). CAM: Causal additive models, high-dimensional order search and penalized regression. *arXiv:1310.1533*.

Cai, R., Zhang, Z., and Hao, Z. (2013). SADA: A general framework to support robust causation discovery. In *Proc. 30th International Conference on Machine Learning (ICML2013)*, pages 208–216.

Chen, Z. and Chan, L. (2013). Causality in linear nonGaussian acyclic models in the presence of latent Gaussian confounders. *Neural Computation*, 25(6):1605–1641.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36:62–83.

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Review of the International Statistical Institute*, 21:2–8.

Dodge, Y. and Rousson, V. (2001). On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician*, 55(1):51–54.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Entner, D. and Hoyer, P. (2010). On causal discovery from time series data using FCI. In *Proc. 5th European Workshop on Probabilistic Graphical Models (PGM2010)*.

Entner, D. and Hoyer, P. O. (2011). Discovering unconfounded causal relationships using linear non-Gaussian models. In *New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science*, volume 6797, pages 181–195.

Entner, D. and Hoyer, P. O. (2012). Estimating a causal order among groups of variables in linear models. In *Proc. 22nd International Conference on Artificial Neural Networks (ICANN2012)*, pages 83–90.

Eriksson, J. and Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11:601–604.

Ferkingsta, E., Lølanda, A., and Wilhelmsen, M. (2011). Causal modeling and inference for electricity markets. *Energy Economics*, 33(3):404–412.

Gao, W. and Yang, H. (2012). Identifying structural VAR model with latent variables using

overcomplete ICA. *Far East Journal of Theoretical Statistics*, 40(1):31–44.

Glymour, C. (2010). What is right with 'Bayes net methods' and what is wrong with 'hunting causes and using them'? *The British Journal for the Philosophy of Science*, 61(1):161–211.

Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Proc. 16th International Conference on Algorithmic Learning Theory (ALT2005)*, pages 63–77.

Henao, R. and Winther, O. (2011). Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, 12:863–905.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time-series via clustering and visualization. *NeuroImage*, 22:1214–1222.

Hirayama, J. and Hyvärinen, A. (2011). Structural equations and divisive normalization for energy-dependent component analysis. In *Advances in Neural Information Processing Systems 23*, pages 1872–1880.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–970.

Hoyer, P. O. and Hyttinen, A. (2009). Bayesian discovery of linear acyclic causal models. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pages 240–248.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G., and Shimizu, S. (2008a). Causal discovery of linear acyclic models with arbitrary distributions. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI2008)*, pages 282–289.

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 689–696.

Hoyer, P. O., Shimizu, S., Kerminen, A., and Palviainen, M. (2008b). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.

Hurley, D., Araki, H., Tamada, Y., Dunmore, B., Sanders, D., Humphreys, S., Affara, M., Imoto, S., Yasuda, K., Tomiyasu, Y., et al. (2012). Gene network inference and visualization tools for biologists: Application to new human transcriptome datasets. *Nucleic Acids Research*, 40(6):2377–2398.

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

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634.

Hyvärinen, A. (2013). Independent component analysis: Recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371:20110534.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley, New York.

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

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregressive model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., and Miyano, S. (2002). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. In *Proc. 1st IEEE Computer Society Bioinformatics Conference*, pages 219–227.

Jutten, C. and Hérault, J. (1991). Blind separation of sources, part I: An adaptive algorithm

based on neuromimetic architecture. *Signal Processing*, 24(1):1–10.

Kadowaki, K., Shimizu, S., and Washio, T. (2013). Estimation of causal structures in longitudinal data using non-Gaussianity. In *Proc. 23rd IEEE International Workshop on Machine Learning for Signal Processing (MLSP2013)*. In press.

Kawahara, Y., Bollen, K., Shimizu, S., and Washio, T. (2010). GroupLiNGAM: Linear non-Gaussian acyclic models for sets of variables. *arXiv:1006.5041*.

Kawahara, Y., Shimizu, S., and Washio, T. (2011). Analyzing relationships among ARMA processes based on non-Gaussianity of external influences. *Neurocomputing*, 4(12–13):2212–2221.

Komatsu, Y., Shimizu, S., and Shimodaira, H. (2010). Assessing statistical reliability of LiNGAM via multiscale bootstrap. In *Proc. 20th International Conference on Artificial Neural Networks (ICANN2010)*, pages 309–314.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI2008)*, pages 366–374.

Lewicki, M. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12(2):337–365.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1):156–166.

Moneta, A., Entner, D., Hoyer, P., and Coad, A. (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75:705–730.

Mooij, J., Janzing, D., Heskes, T., and Schölkopf, B. (2011). Causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24*, pages 639–647.

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

Neyman, J. (1923). *Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle*.

Niyogi, D., Kishtawal, C., Tripathi, S., and Govindaraju, R. S. (2010). Observational evidence that agricultural intensification and land use change may be reducing the Indian summer monsoon rainfall. *Water Resources Research*, 46:W03533.

Ozaki, K. and Ando, J. (2009). Direction of causation between shared and non-shared environmental factors. *Behavior Genetics*, 39(3):321–336.

Ozaki, K., Toyoda, H., Iwama, N., Kubo, S., and Ando, J. (2011). Using non-normal SEM to resolve the ACDE model in the classical twin design. *Behavior Genetics*, 41(2):329–339.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. (2nd ed. 2009).

Pearl, J. and Verma, T. (1991). A theory of inferred causation. In Allen, J., Fikes, R., and Sandewall., E., editors, *Proc. 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 441–452. Morgan Kaufmann, San Mateo, CA.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, 144:864–873.

Peters, J., Janzing, D., and Schölkopf, B. (2011a). Causal inference on discrete data using

additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems 26*.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2011b). Identifiability of causal graphs using functional models. *Proc. 27th Conference on Uncertainty in Artificial Intelligence (UAI2011)*, pages 589–598.

Ramsey, J., Hanson, S., and Glymour, C. (2011). Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simulation study. *NeuroImage*, 58(3):838–848.

Richardson, T. (1996). A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Proc. 12th Conference on Uncertainty in Artificial Intelligence (UAI1996)*, pages 462–469.

Rosenström, T., Jokela, M., Puttonen, S., Hintsanen, M., Pulkki-Råback, L., Viikari, J. S., Raitakari, O. T., and Keltikangas-Järvinen, L. (2012). Pairwise measures of causal direction in the epidemiology of sleep problems and depression. *PloS ONE*, 7(11):e50841.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Schaechtle, U., Stathis, K., Holloway, R., and Bromuri, S. (2013). Multi-dimensional causal discovery. In *Proc. 23rd International Joint Conference on Artificial Intelligence (IJCAI2013)*, pages 1649–1655.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *Proc. 29th International Conference on Machine learning (ICML2012)*, pages 1255–1262.

Shimizu, S. (2012). Joint estimation of linear non-Gaussian acyclic models. *Neurocomputing*, 81:104–107.

Shimizu, S. and Bollen, K. (2013). Bayesian estimation of possible causal direction in the presence of latent confounders using a linear non-Gaussian acyclic structural equation model with individual-specific effects. *arXiv:1310.6778*.

Shimizu, S., Hoyer, P. O., and Hyvärinen, A. (2009). Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, 72:2024–2027.

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

Shimizu, S. and Hyvärinen, A. (2008). Discovery of linear non-Gaussian acyclic models in the presence of latent classes. In *Proc. 14th International Conference on Neural Information Processing (ICONIP2007)*, pages 752–761.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248.

Shimizu, S. and Kano, Y. (2008). Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference*, 138:3483–3491.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proc. 22nd Conference on Uncertainty in Artificial Intelligence (UAI2006)*, pages 437–444.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979.

Skitovitch, W. P. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217–219.

Smith, S. (2012). The future of FMRI connectivity. *NeuroImage*, 62(2):1257–1266.

Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J.,

and Woolrich, M. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891.

Sogawa, Y., Shimizu, S., Shimamura, T., Hyvärinen, A., Washio, T., and Imoto, S. (2011). Estimating exogenous variables in data with more variables than observations. *Neural Networks*, 24(8):875–880.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:67–72.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer Verlag. (2nd ed. MIT Press, 2000).

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proc. 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI1995)*, pages 491–506.

Statnikov, A., Henaff, M., Lytkin, N. I., and Aliferis, C. F. (2012). New methods for separating causes from effects in genomics data. *BMC Genomics*, 13(Suppl 8):S22.

Swanson, N. and Granger, C. (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, pages 357–367.

Takahashi, Y., Ozaki, K., Roberts, B., and Ando, J. (2012). Can low behavioral activation system predict depressive mood?: An application of non-normal structural equation modeling. *Japanese Psychological Research*, 54(2):170–181.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. (2014). ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural Computation*.

Thamvitayakul, K., Shimizu, S., Ueno, T., Washio, T., and Tashiro, T. (2012). Bootstrap confidence intervals in DirectLiNGAM. In *Proc. 2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW2012)*, pages 659–668. IEEE.

Tillman, R. E., Gretton, A., and Spirtes, P. (2010). Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems 22*, pages 1847–1855.

Tillman, R. E. and Spirtes, P. (2011). When causality matters for prediction: Investigating the practical tradeoffs. In *JMLR Workshop and Conference Proceedings, Causality: Objectives and Assessment (Proc. NIPS2008 Workshop on Causality)*, volume 6, pages 373–382.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585.

Zhang, K. and Chan, L.-W. (2006). ICA with sparse connections. In *Proc. 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2006)*, pages 530–537.

Zhang, K. and Hyvärinen, A. (2009a). Causality discovery with additive disturbances: An information-theoretical perspective. In *Proc. European Conference on Machine Learning (ECML2009)*, pages 570–585.

Zhang, K. and Hyvärinen, A. (2009b). On the identifiability of the post-nonlinear causal model. In *Proc. 25th Conference in Uncertainty in Artificial Intelligence (UAI2009)*, pages 647–655.

Zhang, K., Schölkopf, B., and Janzing, D. (2010). Invariant Gaussian process latent variable models and application in causal discovery. In *Proc. 26nd Conference on Uncertainty in Artificial Intelligence (UAI2010)*, pages 717–724.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

(Received August 8 2013, Revised October 1 2013)