Journal of Machine Learning Research 15 (2014) 2629-2652
Submitted 10/13; Revised 4/14; Published 8/14

Bayesian Estimation of Causal Direction in Acyclic Structural Equation Models with Individual-specific Confounder Variables and Non-Gaussian Distributions

Shohei Shimizu
SSHIMIZU@AR.SANKEN.OSAKA-U.AC.JP *The Institute of Scientific and Industrial Research*
Osaka University
Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan

Kenneth Bollen
BOLLEN@UNC.EDU *Department of Sociology, CB 3210*
Hamilton Hall
University of North Carolina
Chapel Hill, NC 27599-3210
U.S.A.

Editor: David Maxwell Chickering

Abstract

Several existing methods have been shown to consistently estimate causal direction assuming linear or some form of nonlinear relationship and no latent confounders. However, the estimation results could be distorted if either assumption is violated. We develop an approach to determining the possible causal direction between two observed variables when latent confounding variables are present . We first propose a new linear non-Gaussian acyclic structural equation model with individual-specific effects that are sometimes the source of confounding. Thus, modeling individual-specific effects as latent variables allows latent confounding to be considered. We then propose an empirical Bayesian approach for estimating possible causal direction using the new model. We demonstrate the effectiveness of our method using artificial and real-world data.

Keywords: structural equation models, Bayesian networks, estimation of causal direction, latent confounding variables, non-Gaussianity

1. Introduction

Aids to uncover the causal structure of variables from observational data are welcomed additions to the field of machine learning (Pearl, 2000; Spirtes et al., 1993). One conventional approach makes use of Bayesian networks (Pearl, 2000; Spirtes et al., 1993). However, these suffer from the identifiability problem. That is, many different causal structures give the same conditional independence between variables, and in many cases one cannot uniquely estimate the underlying causal structure without prior knowledge (Pearl, 2000; Spirtes et al., 1993).

To address these issues, Shimizu et al. (2006) proposed LiNGAM (Linear Non-Gaussian Acyclic Model), a variant of Bayesian networks (Pearl, 2000; Spirtes et al., 1993) and structural equation models (Bollen, 1989) . Unlike conventional Bayesian networks, LiNGAM is a fully identifiable model (Shimizu et al., 2006), and has recently attracted much attention in machine learning (Spirtes et al., 2010; Moneta et al., 2011). If causal relations exist among variables, LiNGAM uses their non-Gaussian distributions to identify the causal structure among the variables. LiNGAM is closely related to independent component analysis (ICA) (Hyvärinen et al., 2001b); the identifiability proof and estimation algorithm are partly based on the ICA theory. The idea of LiNGAM has been extended in many directions, including to nonlinear cases (Hoyer et al., 2009; Lacerda et al., 2008; Hyvärinen et al., 2010; Zhang and Hyvärinen, 2009; Peters et al., 2011a).

Many causal discovery methods including LiNGAM make the strong assumption of no latent confounders (Spirtes and Glymour, 1991; Dodge and Rousson, 2001; Shimizu et al., 2006; Hyvärinen and Smith, 2013; Hoyer et al., 2009; Zhang and Hyvärinen, 2009) . These methods have been used in various application fields

(Ramsey et al., 2014; Rosenström et al., 2012; Smith et al., 2011; Statnikov et al., 2012; Moneta et al., 2013). However, in many areas of empirical science, it is often difficult to accept the estimation results because latent confounders are ignored. In theory, we could take a non-Gaussian approach (Hoyer et al., 2008b) that uses an extension of ICA with more latent variables than observed variables (overcomplete ICA) to formally consider latent confounders in the framework of LiNGAM. Unfortunately, current versions of the overcomplete ICA algorithms are not very computationally reliable since they often suffer from local optima (Entner and Hoyer, 2011).

Thus, in this paper, we propose an alternative Bayesian approach to develop a method that is computationally simple in the sense that no iterative search in the parameter space is required and it is capable of finding the possible causal direction of two observed variables in the presence of latent confounders. We first propose a variant of LiNGAM with individual-specific effects. Individual differences are sometimes the source of confounding (von Eye and Bergman, 2003) . Thus, modeling certain individual-specific effects as latent variables allows a type of latent confounding to be considered. A latent confounding variable is an unobserved variable that exerts a causal influence on more than one observed variables (Hoyer et al., 2008b). The new model is still linear but allows any number of latent confounders. We then present a Bayesian approach for estimating the model by integrating out some of the large number of parameters, which is of the same order as the sample size. Such a Bayesian approach is often used in the field of mixed models (Demidenko, 2004) and multilevel models (Kreft and De Leeuw, 1998), although estimation of causal direction is not a topic studied within it.

Granger causality (Granger, 1969) is another popular method to aid detection of causal direction. His method depends on the temporal ordering of variables whereas our method does not. Therefore, our method can be applied to cases where temporal information is not available, i.e., cross-sectional data, as well as those where it is available, i.e., time-series data.

The remainder of this paper is organized as follows. We first review LiNGAM (Shimizu et al., 2006) and its extension to latent confounder cases (Hoyer et al., 2008b) in Section 2. In Section 3, we propose a new mixed-LiNGAM model, which is a variant of LiNGAM with individual-specific effects. We also propose an empirical Bayesian approach for learning the model. We empirically evaluate the performance of our method using artificial and real-world sociology data in Sections 4 and 5, respectively, and present our conclusions in Section 6.

2. Background

In this section, we first review the linear non-Gaussian structural equation model known as LiNGAM (Shimizu et al., 2006). We then discuss an extension of LiNGAM to cases where latent confounding variables exist (Hoyer et al., 2008b).

In LiNGAM (Shimizu et al., 2006), causal relations between observed variables x_l ($l = 1, \cdots, d$) are modeled as

$$x_l = \mu_l + \sum_{k(m)<k(l)} b_{lm} x_m + e_l, \quad (1)$$

where $k(l)$ is a causal ordering of the variables x_l. The causal orders $k(l)$ ($l = 1, \cdots, d$) are unknown and to be estimated. In this ordering, the variables x_l form a directed acyclic graph (DAG) so that no later variable determines, i.e., has a directed path to, any earlier variable in the

DAG. The variables e_l are latent continuous variables called error variables, μ_l are intercepts or regression constants, and b_{lm} are connection strengths or regression coefficients.

In matrix form, the LiNGAM model in Equation (1) is written as

$$\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{x} + \boldsymbol{e}, \tag{2}$$

where the vector $\boldsymbol{\mu}$ collects constants μ_l, the connection strength matrix $\mathbf{B}$ collects regression coefficients (or connection strengths) b_{lm}, and the vectors $\boldsymbol{x}$ and $\boldsymbol{e}$ collect observed variables x_l and error variables e_l , respectively. The zero/non-zero pattern of b_{lm} corresponds to the absence/existence pattern of directed edges (direct effects). 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. Errors e_l follow non-Gaussian distributions with zero mean and non-zero variance, and are jointly independent . This model without assuming non-Gaussianity distribution is called a fully recursive model in conventional structural equation models (Bollen, 1989) . The non-Gaussianity assumption on e_l enables the identification of a causal ordering $k(l)$ and the coefficients b_{lm} based only on $\boldsymbol{x}$ (Shimizu et al., 2006) , unlike conventional Bayesian networks based on the Gaussianity assumption on e_l (Spirtes et al., 1993). To illustrate the LiNGAM model, the following example is considered, whose corresponding directed acyclic graph is provided in Figure 1:

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

In this example, x_3 is *equal to* error e_3 and is exogenous since it is not affected by either of the other two variables x_1 and x_2. Thus, x_3 is in the first position of such a causal ordering such that $\mathbf{B}$ is strictly lower triangular, x_1 is in the second, and x_2 is 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 have

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

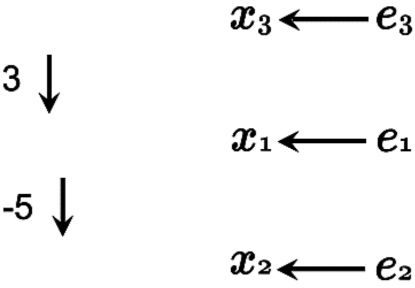


Figure 1: An example graph of LiNGAMs

It can be seen that the resulting connection strength (or regression coefficient) matrix is strictly lower triangular.

Several computationally efficient algorithms for estimating the model have been proposed (Shimizu et al., 2006, 2011; Hyvärinen and Smith, 2013). As with ICA, LiNGAM is identifiable under the assumptions of non-Gaussianity and independence among error variables (Shimizu et al., 2006; Comon, 1994; Eriksson and Koivunen, 2003).[1] However, for the estimation methods to be consistent, additional assumptions, e.g., the existence of their moments or some other statistic, must be made to ensure that the statistics computed in the estimation algorithms exist. The idea of LiNGAM can be generalized to nonlinear cases (Hoyer et al., 2009; Tillman et al., 2010; Zhang and Hyvärinen, 2009; Peters et al., 2011b).

The assumption of independence among e_l means that there is no latent confounding variable (Shimizu et al., 2006). A latent confounding variable is an unobserved variable that contributes to the values of more than one observed variable (Hoyer et al., 2008b). However, in many applications, there often exist latent confounding variables. If such latent confounders are completely ignored, the estimation results can be seriously biased (Pearl, 2000; Spirtes et al., 1993; Bollen, 1989). Therefore, in Hoyer et al. (2008b), LiNGAM with latent confounders, called latent variable LiNGAM, was proposed, and the model can be formulated as follows:

$$x_l = \mu_l + \sum_{k(m)<k(l)} b_{lm} x_m + \sum_{q=1}^{Q} \lambda_{lq} f_q + e_l,$$

where f_q are non-Gaussian individual-specific effects f_q with zero mean and unit variance and λ_{lq} denote the regression coefficients (connection strengths) from f_q to x_l. This model is written in matrix form as follows:

$$\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{x} + \boldsymbol{\Lambda f} + \boldsymbol{e}, \quad (3)$$

where the difference from LiNGAM in Equation

(2) is the existence of a latent confounding variable vector $\boldsymbol{f}$. The vector $\boldsymbol{f}$ collects f_q. The matrix $\boldsymbol{\Lambda}$ collects λ_{lq} and is assumed to

1. Comon (1994) and Eriksson and Koivunen (2003) established the identifiability of ICA based on the characteristic functions of variables. Moments of some variables may not exist, but their characteristic functions always exist.

be of full column rank. Another way to represent latent confounder cases would be to use dependent error variables. Denoting $\boldsymbol{\Lambda f}+\boldsymbol{e}$ in Equation (3) by $\tilde{\boldsymbol{e}}$, we have

$$\begin{aligned}\boldsymbol{x} &= \boldsymbol{\mu}+\mathbf{B}\boldsymbol{x}+\boldsymbol{\Lambda f}+\boldsymbol{e}\\ &= \boldsymbol{\mu}+\mathbf{B}\boldsymbol{x}+\tilde{\boldsymbol{e}},\end{aligned}$$

where $\tilde{e}_i$ are dependent due to the latent confounders f_q . Observed variables that are equal to dependent errors $\tilde{e}_i$ are connected by bi-directed arcs in their graphs. An example graph is given in Figure 4. This representation can be more general since it is easier to extend it to represent nonlinearly dependent errors. In this paper, however, we use the aforementioned representation using independent errors and latent confounders since linear relations of the observed variables, latent confounders, and errors are necessary for our approach.

Without loss of generality, the latent confounders f_q are assumed to be jointly independent since any dependent latent confounders can be remodeled by linear combinations of independent latent variables if the underlying model is linear acyclic and the error variables

are independent (Hoyer et al., 2008b).

considered:

$$\bar{f}_1 =$$

$$\bar{f}_2 =$$

$$
\begin{aligned}
x_1 &= \\
x_2 &= \\
x_3 &= \\
x_4 &=
\end{aligned}
$$

To illustrate this, the following example model is

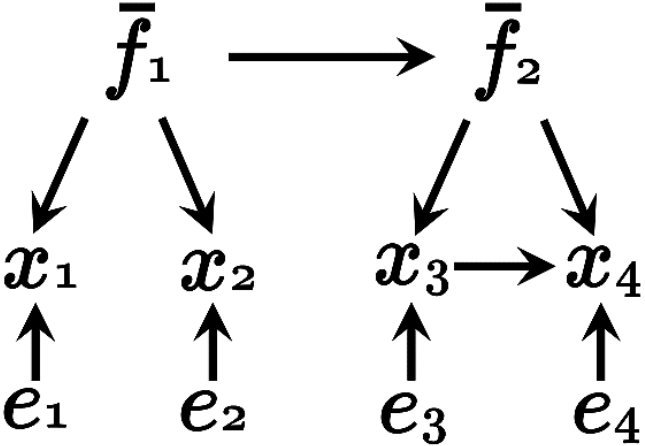
$$
\begin{aligned}
& e_{\bar{f}_1} && (4) \\
& \omega_{21}\bar{f}_1 + e_{\bar{f}_2} && (5) \\
& \lambda_{11}\bar{f}_1 + e_1 \\
& \lambda_{21}\bar{f}_1 + e_2 \\
& \lambda_{32}\bar{f}_2 + e_3 \\
& b_{43}x_3 + \lambda_{42}\bar{f}_2 + e_4,
\end{aligned}
$$

where errors $e_{\bar{f}_1}(=\bar{f}_1)$, $e_{\bar{f}_2}$, and e_1–e_4 are non-Gaussian and independent. The associated graph is shown in Figure 2. The relations of f_1, f_2, and x_1–x_4 are represented by a directed acyclic graph and latent confounders $\bar{f}_1$ and $\bar{f}_2$ are dependent. In matrix form, this example model can be written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ b_{43} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ 0 & \lambda_{32} \\ 0 & \lambda_{42} \end{bmatrix} \begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}.$$

The relations of $\bar{f}_1$ and $\bar{f}_2$ to $e_{\bar{f}_1}$ and $e_{\bar{f}_2}$ in Equations (4)–(5):

$$\begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \omega_{21} & 1 \end{bmatrix} \begin{bmatrix} e_{\bar{f}_1} \\ e_{\bar{f}_2} \end{bmatrix},$$


$\bar{f}_1$
$\bar{f}_2$
x_1
x_2
x_3
x_4
e_1
e_2
e_3
e_4

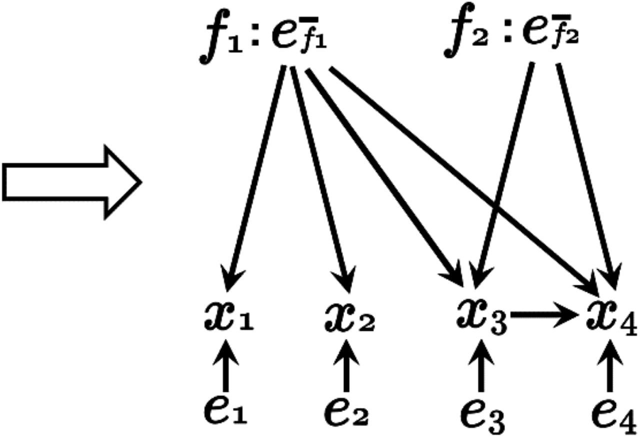


Figure 2: An example graph to illustrate the idea of independent latent confounders.

we obtain

$$
\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ b_{43} & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{\boldsymbol{x}}
$$

$$
+ \underbrace{\begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{32}\omega_{21} & \lambda_{32} \\ \lambda_{42}\omega_{21} & \lambda_{42} \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} e_{\bar{f}_1} \\ e_{\bar{f}_2} \end{bmatrix}}_{\boldsymbol{f}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}}_{\boldsymbol{e}}.
$$

This is a latent variable LiNGAM in Equation (3) taking $f_1 = e_{\bar{f}_1}$ and $f_2 = e_{\bar{f}_2}$ since $e_{\bar{f}_1}$ and $e_{\bar{f}_2}$ are non-Gaussian and independent.

Moreover, the faithfulness of x_l and f_q to the generating graph is assumed. The faithfulness assumption (Spirtes et al., 1993) here means that when multiple causal paths exist from one variable to another, their combined effect does not equal exactly zero (Hoyer et al., 2008b). The faithfulness assumption can be considered to be not very restrictive from the Bayesian viewpoint (Spirtes et al., 1993) since the probability of having exactly the parameter values that do not satisfy faithfulness is zero (Meek, 1995).

In the framework of latent variable LiNGAM, it has been shown (Hoyer et al., 2008b) that the following three models are distinguishable based on observed data,[2] i.e., the three

2. If one or more error variables or latent confounders are Gaussian, it cannot be ensured that Models 3 to 5 will be distinguishable. Hoyer et al. (2008a) considered cases with one or more Gaussian error variables in the context of basic LiNGAM.

different causal structures induce different data distributions:

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

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

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

,

where $\lambda_{1q}\lambda_{2q} \neq 0$ due to the definition of latent confounders, that is, that they contribute to determining the values of more than two variables.

An estimation method based on overcomplete ICA (Lewicki and Sejnowski, 2000) explicitly modeling all the latent confounders f_q was proposed (Hoyer et al., 2008b). However, in current practice, overcomplete ICA estimation algorithms often get stuck in local optima and are not sufficiently reliable (Entner and Hoyer, 2011). A Bayesian approach for estimating the latent variable LiNGAM in Equation (3) has been proposed in Henao and Winther (2011). These previous approaches that explicitly model latent confounders (Hoyer et al., 2008b; Henao and Winther, 2011) need to select the number of latent confounders, and which can be quite large. This could lead to further computational difficulty and statistically unreliable estimates.

In Chen and Chan (2013), a simple approach based on fourth-order cumulants for estimating latent variable LiNGAM was proposed. Their approach does not need to explicitly model the latent confounders, however it requires the latent confounders f_q to be Gaussian. The development of nonlinear methods that incorporate latent confounders is ongoing (Zhang et al., 2010).

None of these latent confounder methods incorporate the individual-specific effects that we model in the next section to consider latent confounders f_q in the latent variable LiNGAM of Equation (3).

3. Linear Non-Gaussian Acyclic Structural Equation Model with Individual-specific Effects

In this section, we propose a new Bayesian method for learning the possible causal direction of two observed variables in the presence of latent confounding variables, assuming that the causal relations are acyclic, i.e., there is not a feedback relation.

3.1 Model

The LiNGAM (Shimizu et al., 2006) for observation i can be described as follows:

$$x_l^{(i)} \;=\; \mu_l + \sum_{k(m)<k(l)} b_{lm} x_m^{(i)} + e_l^{(i)}.$$

The random variables $e_l^{(i)}$ are non-Gaussian and independent. The distributions of $e_l^{(i)}$ $(i = 1, \cdots, n)$ are commonly assumed to be identical[3] for every l. A linear non-Gaussian acyclic structural equation model with individual-specific effects for observation i is formulated as follows:

$$x_l^{(i)} \;=\; \mu_l + \tilde{\mu}_l^{(i)} + \sum_{k(m)<k(l)} b_{lm} x_m^{(i)} + e_l^{(i)},$$

where the difference from LiNGAM is the existence of individual-specific effects $\tilde{\mu}_l^{(i)}$. The parameters $\tilde{\mu}_l^{(i)}$ are independent of $e_l^{(i)}$ and are correlated with $x_l^{(i)}$ through the structural equations in our Bayesian approach, introduced below. This means that the observations are generated from the identifiable LiNGAM, possibly with different parameter values of the means $\mu_l + \tilde{\mu}_l^{(i)}$. We call this a mixed-LiNGAM, named after mixed models (Demidenko, 2004), as it has effects μ_l and b_{lm} that are common to all the observations and individualspecific effects $\tilde{\mu}_l^{(i)}$. We note that causal orderings of variables $k(l)$ $(l = 1, \cdots, d)$ are identical for all the observations in the sample.

To use a Bayesian approach for estimating the mixed-LiNGAM, we need to model the distributions of error variables e_l and prior distributions of the parameters including individual-specific effects $\tilde{\mu}_l^{(i)}$, unlike previous LiNGAM methods (Shimizu et al., 2006; Hoyer et al., 2008b). These individual-specific effects, whose number is of the same order as the sample size, are integrated out in the Bayesian method developed in Section 3.2, assuming an informative prior for them similar to the estimation of conventional mixed models (Demidenko, 2004). More details on the distributions of error variables and prior distributions of parameters are given in Section 3.2. These distributional assumptions were implied to be robust to some extent to their violations, at least in the artificial data experiments of Section 4.

We now relate the mixed-LiNGAM model above with the latent variable LiNGAM (Hoyer et al., 2008b). The latent variable LiNGAM in Equation (3) for observation i is written as follows:

$$x_l^{(i)} = \mu_l + \sum_{k(m)<k(l)} b_{lm} x_m^{(i)} + \underbrace{\sum_{q=1}^{Q} \lambda_{lq} f_q^{(i)}}_{\tilde{\mu}_l^{(i)}} + e_l^{(i)}.$$

This is a mixed-LiNGAM taking $\tilde{\mu}_l^{(i)} = \sum_{q=1}^{Q} \lambda_{lq} f_q^{(i)}$. In contrast to the previous approaches for latent variable LiNGAM (Hoyer et al., 2008b; Henao and Winther, 2011), we do not explicitly model the latent confounders f_q and rather simply include their sums $\tilde{\mu}_l^{(i)} = \sum_{q=1}^{Q} \lambda_{lq} f_q^{(i)}$ in our model as its parameters since our main interest lies in estimation of the causal relation of observed variables x_l and not in the estimation of their relations with latent confounders f_q. Our method does not estimate λ_{lq} or the number of latent confounders Q.

---

3 . Relaxing this identically distributed assumption would lead to more general modeling of individual differences, however, this goes beyond the scope of the paper.

3.2 Estimation of Possible Causal Direction

We apply a Bayesian approach to estimate the possible causal direction of two observed variables using the mixed-LiNGAM proposed above. We compare the following two mixed-LiNGAM models with opposite possible directions of causation. Model 1 is

$$\begin{aligned} x_1^{(i)} &= \mu_1 + \tilde{\mu}_1^{(i)} + e_1^{(i)} \\ x_2^{(i)} &= \mu_2 + \tilde{\mu}_2^{(i)} + b_{21} x_1^{(i)} + e_2^{(i)}, \end{aligned}$$

where b_{21} is non-zero. In Model 1, x_2 does not cause x_1. The second model, Model 2, is

$$\begin{aligned} x_1^{(i)} &= \mu_1 + \tilde{\mu}_1^{(i)} + b_{12} x_2^{(i)} + e_1^{(i)} \\ x_2^{(i)} &= \mu_2 + \tilde{\mu}_2^{(i)} + e_2^{(i)}, \end{aligned}$$

where b_{12} is non-zero. In Model 2, x_1 does not cause x_2. The two models have the same number of parameters, but opposite possible directions of causation.

Once the possible causal direction is estimated, one can see if the common causal coefficient (connection strength) b_{21} or b_{12} is likely to be zero by examining its posterior distribution.[4] We focus here on estimating the possible direction of causation as in many previous works (Dodge and Rousson, 2001; Hoyer et al., 2009; Zhang and Hyvärinen, 2009; Chen and Chan, 2013; Hyvärinen and Smith, 2013), and do not go to the computation of the posterior distribution[5] since estimation of the possible causal direction of two observed variables in the presence of latent confounders has been a very challenging problem in causal inference and is the main topic of this paper.

We apply standard Bayesian model selection techniques to help assess the causal direction of x_1 and x_2. We use the log-marginal likelihood for comparing the two models. The model with the larger log-marginal likelihood is regarded as the closest to the true model (Kass and Raftery, 1995).

Let $\mathcal{D}$ be the observed data set $[\boldsymbol{x}^{(1)^T}, \cdots, \boldsymbol{x}^{(n)^T}]^T$, where $\boldsymbol{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}]^T$. Denote Models 1 and 2 by M_1 and M_2. The log-marginal likelihoods of M_1 and M_2 are

$$\begin{aligned}\log\{p(M_r|\mathcal{D})\} &= \log\{p(\mathcal{D}|M_r)p(M_r)/p(\mathcal{D})\} \\ &= \log\{p(\mathcal{D}|M_r)\} + \log\{p(M_r)\} - \log p(\mathcal{D}) = \\ &\log\{\int p(\mathcal{D}|\boldsymbol{\theta}_r, M_r)p(\boldsymbol{\theta}_r|M_r, \boldsymbol{\eta}_r)d\boldsymbol{\theta}_r\} \\ &+ \log p(M_r) - \log p(\mathcal{D}) \ (r = 1, 2),\end{aligned}$$

where $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ are the hyper-parameter vectors regarding the distributions of the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Since the last term $\log p(\mathcal{D})$ is constant with respect to M_r, we can drop it. To select suitable values for these hyper-parameters, we take an ordinary empirical Bayesian approach. First, we compute the log-marginal likelihood for every combination

4. Chickering and Pearl (1996) considered a discrete variable model with *known* possible causal direction

and proposed a Bayesian approach for computing the posterior distributions of causal effects in the presence of latent confounders.

5. Point estimates of the parameters including the common causal connection strengths b_{12} and b_{21} can be obtained by taking their posterior means based on their posterior distributions, for example.

of the two models M_r and a number of candidate hyper-parameter values of $\boldsymbol{\eta}_r$. Next, we take the model and hyper-parameter values that give the largest log-marginal likelihood, and finally estimate that the model with the largest log-marginal likelihood is better than the other model.

In basic LiNGAM (Shimizu et al., 2006), we have (Hyvärinen et al., 2010; Hoyer and Hyttinen, 2009)

$$p(\boldsymbol{x}) = \prod_l p_{e_l}\left(x_l - \mu_l - \sum_{k(m)<k(l)} b_{lm} x_m\right).$$

Thus, in the same manner, the likelihoods under mixed-LiNGAM $p(\mathcal{D}|\boldsymbol{\theta}_r, M_r)$ $(r = 1, 2)$ are given by

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}_r, M_r) &= \Pi_{i=1}^{n}\, p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}_r, M_r) \\
&= \begin{cases} \Pi_{i=1}^{n}\, p_{e_1^{(i)}}(x_1^{(i)} - \mu_1 - \tilde{\mu}_1^{(i)}|\boldsymbol{\theta}_1, M_1) \\ \times\, p_{e_2^{(i)}}(x_2^{(i)} - \mu_2 - \tilde{\mu}_2^{(i)} - b_{21}x_1^{(i)}|\boldsymbol{\theta}_1, M_1) \ \text{ for } M_1 \\ \Pi_{i=1}^{n}\, p_{e_1^{(i)}}(x_1^{(i)} - \mu_1 - \tilde{\mu}_1^{(i)} - b_{12}x_2^{(i)}|\boldsymbol{\theta}_2, M_2) \\ \times\, p_{e_2^{(i)}}(x_2^{(i)} - \mu_2 - \tilde{\mu}_2^{(i)}|\boldsymbol{\theta}_2, M_2) \text{ for } M_2 \end{cases}.
\end{aligned}
$$

We model the parameters and their prior distributions as follows.[6] The prior probabilities of M_1 and M_2 are uniform:

$$p(M_1) = p(M_2).$$

The distributions of the error variables $e_1^{(i)}$ and $e_2^{(i)}$ are modeled by Laplace distributions with zero mean and variances of $\mathrm{var}(e_1^{(i)}) = h_1^2$ and $\mathrm{var}(e_2^{(i)}) = h_2^2$ as follows:

$$
\begin{aligned}
p_{e_1^{(i)}} &= Laplace(0, |h_1|/\sqrt{2}) \\
p_{e_2^{(i)}} &= Laplace(0, |h_2|/\sqrt{2}).
\end{aligned}
$$

Here, we simply use a symmetric super-Gaussian distribution, i.e., the Laplace distribution, to model $p_{e_1^{(i)}}$ and $p_{e_2^{(i)}}$, as suggested in Hyvärinen and Smith (2013). Such super-Gaussian distributions have been reported to often work

well in non-Gaussian estimation methods including independent component analysis and LiNGAM (Hyvärinen et al., 2001b; Hyvärinen and Smith, 2013). In some cases, a wider class of non-Gaussian distributions might provide a better model for $p_{e_1^{(i)}}$ and $p_{e_2^{(i)}}$, *e.g.*, the generalized Gaussian family (Hyvärinen et al., 2001b), a finite mixture of Gaussians, or an exponential family distribution combining the Gaussian and Laplace distributions (Hoyer and Hyttinen, 2009)
.

The parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are written as follows:

$$\boldsymbol{\theta}_1 = [\mu_l, b_{21}, h_l, \tilde{\mu}_l^{(i)}]^T \quad (l = 1, 2; i = 1, \cdots, n)$$

$$\boldsymbol{\theta}_2 = [\mu_l, b_{12}, h_l, \tilde{\mu}_l^{(i)}]^T \quad (l = 1, 2; i = 1, \cdots, n).$$

6. This is an example. The modeling method could depend on the domain knowledge.

The prior distributions of common effects are Gaussian as follows:

$$\begin{aligned}
\mu_1 &\sim N(0, \tau_{\mu_1}^{cmmn}) \\
\mu_2 &\sim N(0, \tau_{\mu_2}^{cmmn}) \\
b_{12} &\sim N(0, \tau_{b_{12}}^{cmmn}) \\
b_{21} &\sim N(0, \tau_{b_{21}}^{cmmn}) \\
h_1 &\sim N(0, \tau_{h_1}^{cmmn}) \\
h_2 &\sim N(0, \tau_{h_2}^{cmmn}),
\end{aligned}$$

where $\tau_{\mu_1}^{cmmn}$, $\tau_{\mu_2}^{cmmn}$, $\tau_{b_{12}}^{cmmn}$, $\tau_{b_{21}}^{cmmn}$, $\tau_{h_1}^{cmmn}$ and $\tau_{h_2}^{cmmn}$ are constants.

Generally speaking, we could use various informative prior distributions for the individual-specific effects and then compare candidate priors using the standard model selection approach based on the marginal likelihoods. Below we provide two examples.

If the data is generated from a latent variable LiNGAM, a special case of mixed-LiNGAM, as shown in Section 3.1, the individual-specific effects are the sums of many non-Gaussian independent latent confounders f_q and are dependent. The central limit theorem states that the sum of independent variables becomes increasingly close to the Gaussian (Billingsley, 1986). Therefore, in many cases, it could be practical to approximate the non-Gaussian distribution of a variable that is the sum of many non-Gaussian and independent variables by a bell-shaped curve distribution (Sogawa et al., 2011; Chen and Chan, 2013). This motivates us to model the prior distribution of individual-specific effects by the multivariate t-distribution as follows:

$$\begin{bmatrix} \tilde{\mu}_1^{(i)} \\ \tilde{\mu}_2^{(i)} \end{bmatrix} =$$

$$\mathrm{diag}\left(\left[\sqrt{\tau_1^{indvdl}}, \sqrt{\tau_2^{indvdl}}\right]^T\right)\mathbf{C}^{-1/2}\boldsymbol{u},$$

(6)

where τ_1^{indvdl} and τ_2^{indvdl} are constants, $\boldsymbol{u} \sim$

$t_\nu(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Sigma} = [\sigma_{ab}]$ is a symmetric scale matrix whose diagonal elements are 1s. A random variable vector $\boldsymbol{u}$ that follows the multivariate t-distribution $t_\nu(\mathbf{0}, \mathbf{\Sigma})$ can be created by $\frac{\boldsymbol{y}}{\sqrt{v/\nu}}$, where $\boldsymbol{y}$ follows the Gaussian distribution $N(\mathbf{0}, \mathbf{\Sigma})$, v follows the chi-squared distribution with ν degrees of freedom, and $\boldsymbol{y}$ and v are statistically independent (Kotz and Nadarajah, 2004). Note that u_i have energy correlations (Hyvärinen et al., 2001a), i.e., correlations of squares $\mathrm{cov}(u_i^2, u_j^2) > 0$ due to the common variable v. $\mathbf{C}$ is a diagonal matrix whose diagonal elements give the variance of elements of $\boldsymbol{u}$, i.e., $\mathbf{C} = \frac{\nu}{\nu-2}\mathrm{diag}(\mathbf{\Sigma})$ for $\nu > 2$. The degree of freedom ν is here taken to be six. The kurtosis of the univariate Student's t-distribution with six degrees of freedom is three, the same as that of the Laplace distribution.

The hyper-parameter vectors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are

$$\boldsymbol{\eta}_l = [\tau_{\mu_1}^{cmmn}, \tau_{\mu_2}^{cmmn}, \tau_{b_{12}}^{cmmn}, \tau_{b_{21}}^{cmmn}, \tau_{h_1}^{cmmn}, \tau_{h_2}^{cmmn}, \tau_1^{indvdl}, \tau_2^{indvdl}, \sigma_{21}]^T \quad (l = 1, 2).$$

We want to take the constants $\tau_{\mu_1}^{cmmn}$, $\tau_{\mu_2}^{cmmn}$, $\tau_{b_{12}}^{cmmn}$, $\tau_{b_{21}}^{cmmn}$, $\tau_{h_1}^{cmmn}$ and $\tau_{h_2}^{cmmn}$ to be sufficiently large so that the priors for the common effects are not very informative. It depends on the scales of variables when these constants are sufficiently large. In the experiments in Sections 4–5, we set $\tau_{\mu_1}^{cmmn} = \tau_{b_{12}}^{cmmn} = \tau_{h_1}^{cmmn} = 10^2 \times \widehat{\mathrm{var}}\,(x_1)$ and $\tau_{\mu_2}^{cmmn} =$

$\tau_{b_{21}}^{cmmn} = \tau_{h_2}^{cmmn} = 10^2 \times \widehat{\mathrm{var}}\ (x_2)$ so that they reflect the scales of the corresponding variables.

Moreover, we take an empirical Bayesian approach for the individual-specific effects. We test τ_l^{indvdl} $= 0, 0.2^2 \times \widehat{\mathrm{var}}(x_l), \ldots, 0.8^2 \times \widehat{\mathrm{var}}(x_l), 1.0^2 \times \widehat{\mathrm{var}}(x_l)$ $(l = 1, 2)$. That is, we uniformly vary the hyper-parameter value from that with no individual-specific effects, i.e., 0, to a larger value, i.e., $1.0^2 \times \widehat{\mathrm{var}}(x_l)$, which implies very large individual differences. Further, we test $\sigma_{12} = 0$, ± 0.3, ± 0.5, ± 0.7, ± 0.9, i.e., the value with zero correlation and larger values with stronger correlations. This means that we test uncorrelated individual-specific effects as well as correlated ones. We take the ordinary Monte Carlo sampling approach to compute the log-marginal likelihoods with 1000 samples for the parameter vectors $\boldsymbol{\theta}_r$ $(r = 1, 2)$.

The assumptions for our model are

summarized in Table 1. Generally speaking, if the actual probability density function of individual-specific effects is unimodal and most often provides zero or very small absolute values and with few large values, i.e., many of the individual-specific effects are close to zero and many individuals have similar intercepts, the estimation is likely to work. If the individuals have very different intercepts, the estimation will not work very well.

An alternative way of modeling the prior distribution of individual-specific effects would be to use the multivariate Gaussian distribution as follows:

$$\begin{bmatrix} \tilde{\mu}_1^{(i)} \\ \tilde{\mu}_2^{(i)} \end{bmatrix} = \operatorname{diag}\left(\left[\sqrt{\tau_1^{indvdl}}, \sqrt{\tau_2^{indvdl}} \right]^T \right) \boldsymbol{z},$$

where τ_1^{indvdl} and τ_2^{indvdl} are constants, $\boldsymbol{z} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Sigma} = [\sigma_{ab}]$ is a symmetric scale matrix whose diagonal elements are 1s. This Gaussian prior would be effective if the Gaussian approximation based on the central limit theorem works well, although a non-Gaussian prior would be more consistent with the non-Gaussian latent variable LiNGAM in Equation (3) . Gaussian individual-specific effects or latent confounders would not lead to losing the identifiability (Chen and Chan, 2013) since each observation still is generated by the identifiable non-Gaussian LiNGAM. However, if errors are Gaussian, there is no guarantee that our method can find correct possible causal direction. We could detect their Gaussianity by comparing our mixed-LiNGAM models with Gaussian error models based on their log-marginal likelihoods. If the errors are actually Gaussian or close to be Gaussian, Gaussian error models would provide larger log-marginal likelihoods. This would detect situations where our approach cannot find causal direction.

4. Experiments on Artificial Data

We compared our method with seven methods for estimating the possible causal direction between two variables: i) LvLiNGAM[7] (Hoyer et al., 2008b); ii) SLIM[8](Henao and Winther, 2011) iii) LiNGAM-GC-UK (Chen and Chan, 2013); iv) ICA-LiNGAM[9] (Shimizu et al., 2006); v) DirectLiNGAM[10] (Shimizu et al., 2011); vi) Pairwise LiNGAM[11] (Hyvärinen

---

7. The code is available at `http://www.cs.helsinki.fi/u/phoyer/code/lvlingam.tar.gz`.
8. The code is available at `http://cogsys.imm.dtu.dk/slim/`.
9. The code is available at `http://www.cs.helsinki.fi/group/neuroinf/lingam/lingam.tar.gz`.
10. The code is available at `http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/Dlingamcode.html`.
11. The code is available at `http://www.cs.helsinki.fi/u/ahyvarin/code/pwcausal/`.

Model: $x_l^{(i)} = \mu_l + \tilde{\mu}_l^{(i)} + \sum_{k(m)<k(l)} b_{lm} x_m^{(i)} + e_l^{(i)}$ $(l, m = 1, 2;\ l \neq m)$, where b_{lm} are non-zero.

$e_l^{(i)}$ $(l = 1, 2; i = 1, \cdots, n)$ are i.i.d..

e_l $(l = 1, 2)$ are mutually independent.

e_l $(l = 1, 2)$ follow Laplace distributions with zero mean and standard deviations $|h_l|$.

Prior distributions:
μ_l, b_{lm} and h_l $(l = 1, 2; m = 1, 2; l \neq m)$ follow Gaussian distributions with zero mean and variance $\tau_{\mu_l}^{cmmn}$, $\tau_{b_{lm}}^{cmmn}$ and $\tau_{h_l}^{cmmn}$.
$\tilde{\mu}_l^{(i)}$ $(l = 1, 2; i = 1, \cdots, n)$ are the sum of latent confounders $f_q^{(i)}$: $\sum_{q=1}^{Q} \lambda_{lq} f_q^{(i)}$ and are independent of $e_l^{(i)}$.
$\tilde{\mu}_l^{(i)}$ $(l = 1, 2; i = 1, \cdots, n)$ are i.i.d..
μ_l $(l = 1, 2)$ follow multivariate t-distributions with ν degrees of freedom, zero mean, variances τ_l^{indvdl} and correlation σ_{12} (here, $\nu = 6$).

Hyper-parameters:
$\tau_{\mu_l}^{cmmn}$, $\tau_{b_{lm}}^{cmmn}$ and $\tau_{h_l}^{cmmn}$ $(l = 1, 2; m = 1, 2; l \neq m)$ are set to be large values so that the priors are not very informative.
τ_l^{indvdl} $(l = 1, 2)$ are uniformly varied from zero to large values.
σ_{12} are uniformly varied in the interval between -0.9 and 0.9.

Table 1: Summary of the assumptions for our mixed-LiNGAM model

and Smith, 2013); vii) Post-nonlinear causal model (PNL) [12] (Zhang and Hyvärinen, 2009). Their assumptions are summarized in Table 2. The first seven methods assume linearity, and the eighth allows a very wide variety of nonlinear relations. The last four methods assume that there are no latent confounders. We tested the prior t- and Gaussian distributions for individual-specific effects in our approach. LvLiNGAM and SLIM require to specify the number of latent confounders. We tested 1 and 4 latent confounder(s) for LvLiNGAM since its current implementation cannot handle more than four latent confounders, whereas we tested 1, 4 and 10 latent confounders(s) for SLIM. LiNGAM-GC-UK (Chen and Chan, 2013) assumes that errors are simultaneously super-Gaussian or sub-Gaussian and that latent confounders are Gaussian.

	Functional form?	Latent confounders allowed?	Number of latent confounders necessary to be specified?	Iterative search in the parameter space required?	Distributional assumptions necessary?
Our approach	Linear	Yes	No	No	Yes
LvLiNGAM	Linear	Yes	Yes	Yes	No[13]
SLIM	Linear	Yes	Yes	No	Yes
LiNGAM-GC-UK	Linear	Yes	No	No	Yes
ICA-LiNGAM	Linear	No	N/A	Yes	No
DirectLiNGAM	Linear	No	N/A	No	No
Pairwise LiNGAM	Linear	No	N/A	No	No
PNL	Nonlinear	No	N/A	Yes	No

Table 2: Summary of the assumptions of eight methods

12. The code is available at `http://webdav.tuebingen.mpg.de/causality/CauseOrEffect_NICA.rar`. 13. Their current implementation of LvLiNGAM in Footnote 7 assumes a non-Gaussian distribution, which

is a mixture of two Gaussian distributions.

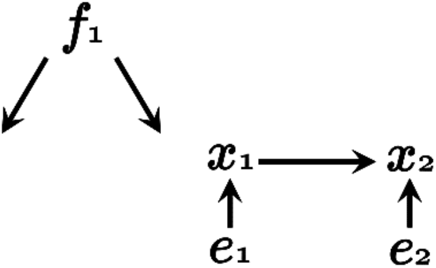


Figure 3: The associated graph of the model used to generate artificial data when the number of latent confounders $Q = 1$.

We generated data using the following latent variable LiNGAM with Q latent confounding variables, which is a mixed-LiNGAM:

$$x_1^{(i)} = \mu_1 + \sum^{Q} \lambda_{1q} f_q^{(i)} + e_1^{(i)}$$

$$q=1$$

$$x_2^{(i)} = \mu_2 + b_{21}x_1^{(i)} + \sum_{q=1}^{Q} \lambda_{2q} f_q^{(i)} + e_2^{(i)},$$

where μ_1 and μ_2 were randomly generated from $N(0,\ 1)$, and b_{21}, λ_{1q}, λ_{2q} were randomly generated from the interval $(-1.5, -0.5) \cup (0.5, 1.5)$. We tested various numbers of latent confounders $Q = 0, 1, 6, 12$. The zero values indicate that there are no latent confounders. An example graph used to generate artificial data is given in Figure 3.

The distributions of the error variables e_1, e_2, and latent confounders f_q were identical for all observations. The distributions of the error variables e_1, e_2, and latent confounders f_q were randomly selected from the 18 non-Gaussian distributions used in Bach and Jordan (2002) to see if the Laplace distribution assumption on error variables and t- or Gaussian distribution assumption on individual-specific effects in our method were robust to different non-Gaussian distributions. These include symmetric/non-symmetric distributions, super-Gaussian/sub-Gaussian distributions, and strongly/weakly non-Gaussian distributions. The variances of e_1 and e_2 were randomly selected from the interval $(0.5^2, 1.5^2)$. The variances of f_q were 1s.

We permuted the variables according to a random ordering and inputted them to the eight estimation methods. We conducted 100 trials, with sample sizes of 50, 100, and 200. For the data with the number of latent confounders $Q = 0$, all the methods should find the correct causal direction for large enough sample sizes, as there were no latent confounders, which here means no individual-specific effects. The last four comparative methods should find the data with the number of latent confounders $Q = 1, 6, 12$ very difficult to analyze, because, unlike the other approaches, they assume no latent confounders.

To evaluate the performance of the algorithms, we counted the number of successful discoveries of possible causal direction and estimated their standard errors.

Looking at Table 3 as a whole there are several general observations that we can make. First though none of the procedures is infallible, several of them do quite well in that they choose the correct causal direction about 90% of the time. Second, overall our approach is the most successful across the conditions of the simulation. Specifically, in all but the cases of no confounding variables, one or both of our approaches have the highest percentages of success. In the situation of no confounding variables, ICA-LiNGAM, DirectLiNGAM, and Pairwise LINGAM have higher success percentages than our procedures. These generalizations need qualifications in that there are sampling errors that affect the estimates. Formal tests of significance across all conditions would be complicated. It would require taking account of multiple testing and the dependencies of the simulated samples under the same sample size and number of confounders.

However, the standard errors of the estimated percentages serve to caution the reader not to judge the percentages alone without recognizing sampling variability. For instance, when there are no confounders and a sample size of 50, the ICA-LiNGAM procedure appears best with 93% success, but the success percentages of our two approaches fall within two standard errors of the 93% estimate. Alternatively, in the rows with 6 confounders and sample size 50 our approach with 88% success and a standard error of 3.25 appears sufficiently far from the success percentages of the other methods besides ours to make sampling fluctuations an unlikely explanation. In sum, taking all the evidence together, our approaches performed quite well and deserve further investigation under additional simulation conditions.

Table 4 shows the average computational times. The computational complexity of the current implementation of our methods is clearly larger than that of the other linear methods ICA-LiNGAM, DirectLiNGAM, Pairwise LiNGAM, LvLiNGAM with 1 latent confounder, SLIM and LiNGAM-GC-UK and comparable to LvLiNGAM with 4 latent confounders and the nonlinear method PNL.

The MATLAB code for performing these experiments is available on our website.[14]

5. An Experiment on Real-world Data

We analyzed the General Social Survey data set, taken from a sociological data repository (`http://www.norc.org/GSS+Website/`). The data consisted of six observed variables: x_1: prestige of father's occupation, x_2 : son's income, x_3: father's education, x_4: prestige of son's occupation, x_5: son's education, and x_6: number of siblings.[15] The sample selection was conducted based on the following criteria: i) non-farm background; ii) ages 35–44; iii) white; iv) male; v) in the labor force at the time of the survey; vi) not missing data for any of the covariates; and vii) data taken from 1972–2006. The sample size was 1380.

The possible directions were determined

based on the domain knowledge in Duncan et al. (1972), shown in Figure 4. Note that there is no direct causal link from x_1, x_3, and x_6 to x_2 in the figure, however it is expected that each of these variables has non-zero total causal effects on x_2 given their indirect effects on x_2. The causal relations of x_1, x_3, and x_6 usually are not modeled in the literature since there are many other determinants of these three exogenous observed variables that are not part of the model. However, the possible

---

14. The URL is `http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/mixedlingamcode.html`.

15. Although x_6 is discrete, it can be considered as continuous because it is an ordinal scale with many points.

	Sample size	50	100	200
Number of latent confounders $Q = 0$:				
Our approach (t-distributed individual-specific effects)		88		

(3.25) 91 (2.86) 86 (3.47) Our approach (Gaussian individual-specific effects) | 91 (2.86) 87 (3.36) 91 (2.86) LvLiNGAM (1 latent confounder) |
73 (4.44) 83 (3.76) 83 (3.76) LvLiNGAM (4 latent confounders) | 52 (5.00) 68 (4.66) 66 (4.74) SLIM (1 latent confounder) |
29 (4.54) 30 (4.58) 25 (4.33) SLIM (4 latent confounders) | 34 (4.74) 31 (4.62) 36 (4.80) SLIM (10 latent confounders) | 30 (4.58) 29 (4.54) 30 (4.58) LiNGAM-GC-UK
| 33 (4.70) 28 (4.49) 35 (4.77) ICA-LiNGAM
93 (2.55) 93 (2.55) 96 (1.96) DirectLiNGAM
87 (3.36) 95 (2.18) 97 (1.71) Pairwise LiNGAM
89 (3.13) 95 (2.18) 95 (2.18) Post-nonlinear causal model
74 (4.39) 71 (4.54) 75 (4.33)

Number of latent confounders $Q = 1$: |

Our approach (t-distributed individual-specific effects) | 83 (3.76) 80 (4.00) 80 (4.00) Our approach (Gaussian individual-specific effects) | 79 (4.07) 87 (3.36) 69 (4.62) LvLiNGAM (1 latent confounder) |
66 (4.74) 71 (4.54) 73 (4.44) LvLiNGAM (4 latent confounders) | 63 (4.83) 58 (4.94) 67 (4.70) SLIM (1 latent confounder) |
40 (4.90) 47 (4.99) 25 (4.33) SLIM (4 latent confounders) | 40 (4.90) 34 (4.74) 44 (4.96) SLIM (10 latent confounders) | 47 (4.99) 39 (4.88) 41 (4.92) LiNGAM-GC-UK
| 24 (4.27) 32 (4.66) 32 (4.66) ICA-LiNGAM
74 (4.39) 71 (4.54) 67 (4.70) DirectLiNGAM
48 (5.00) 52 (5.00) 54 (4.98) Pairwise LiNGAM
54 (4.98) 58 (4.94) 61 (4.88) Post-nonlinear causal model
55 (4.97) 58 (4.94) 57 (4.95)

Number of latent confounders $Q = 6$: |

Our approach (t-distributed individual-specific effects)	88 (3.25)	81 (3.92)	87 (3.36)
Our approach (Gaussian individual-specific effects)	84 (3.67)	85 (3.57)	87 (3.36)
LvLiNGAM (1 latent confounder)	58 (4.94)	70 (4.58)	70 (4.58)
LvLiNGAM (4 latent confounders)	64 (4.80)	61 (4.88)	63 (4.83)
SLIM (1 latent confounder)	50 (5.00)	63 (4.83)	47 (4.99)
SLIM (4 latent confounders)	45 (4.97)	47 (4.99)	43 (4.95)
SLIM (10 latent confounders)	58 (4.94)	48 (5.00)	58 (4.94)
LiNGAM-GC-UK	29 (4.54)	28 (4.49)	21 (4.07)
ICA-LiNGAM	74 (4.39)	72 (4.49)	47 (4.99)
DirectLiNGAM	37 (4.83)	48 (5.00)	39 (4.88)
Pairwise LiNGAM	48 (5.00)	51 (5.00)	37 (4.83)
Post-nonlinear causal model	55 (4.97)	42 (4.94)	46 (4.98)
Number of latent confounders $Q = 12$:			
Our approach (t-distributed individual-specific effects)	88 (3.25)	86 (3.47)	89 (3.13)
Our approach (Gaussian individual-specific effects)	91 (2.86)	89 (3.13)	91 (2.86)
LvLiNGAM (1 latent confounder)	52 (5.00)	55 (4.97)	65 (4.77)
LvLiNGAM (4 latent confounders)	65 (4.77)	58 (4.94)	64 (4.80)
SLIM (1 latent confounder)	51 (5.00)	55 (4.97)	60 (4.90)
SLIM (4 latent confounders)	45 (4.97)	51 (5.00)	63 (4.83)
SLIM (10 latent confounders)	61 (4.88)	54 (4.98)	54 (4.98)
LiNGAM-GC-UK	21 (4.07)	25 (4.33)	29 (4.54)
ICA-LiNGAM	68 (4.66)	72 (4.49)	72 (4.49)
DirectLiNGAM	37 (4.83)	39 (4.88)	38 (4.85)
Pairwise LiNGAM	56 (4.96)	42 (4.94)	43 (4.95)
Post-nonlinear causal model	51 (5.00)	43 (4.95)	46 (4.98)

Largest numbers of successful discoveries were underlined. Standard errors are shown in parentheses, which are computed assuming that the number of successes follow a binomial distribution.

Table 3: Number of successful discoveries (100 trials)

	Sample size		
	50	100	200
Number of latent confounders $Q = 0$			
Our approach (t-distributed individual-specific effects)	27.20	56.93	141.84
Our approach (Gaussian individual-specific effects)	35.48	69.59	117.10
LvLiNGAM (1 latent confounder)	2.41	2.55	9.91
LvLiNGAM (4 latent confounders)	22.25	30.12	87.96
SLIM (1 latent confounder)	5.89	6.25	6.81
SLIM (4 latent confounders)	7.60	8.14	9.13
SLIM (10 latent confounders)	10.88	12.02	13.96
LiNGAM-GC-UK	0.00	0.00	0.00
ICA-LiNGAM	0.04	0.03	0.02
DirectLiNGAM	0.00	0.01	0.01
Pairwise LiNGAM	0.00	0.00	0.00
Post-nonlinear causal model	19.59	27.68	57.37
Number of latent confounders $Q = 1$:			
Our approach (t-distributed individual-specific effects)	35.87	65.55	131.25
Our approach (Gaussian individual-specific effects)	37.12	75.11	114.37

LvLiNGAM (1 latent confounder)
LvLiNGAM (4 latent confounders)
SLIM (1 latent confounder)
SLIM (4 latent confounders)

	2.40	2.53	13.93
	21.50	29.50	92.19
	5.88	6.01	6.69
	7.59	8.19	8.96
SLIM (10 latent confounders)	10.96	11.79	13.68
LiNGAM-GC-UK	0.00	0.00	0.00
ICA-LiNGAM	0.03	0.03	0.05
DirectLiNGAM	0.01	0.01	0.01
Pairwise LiNGAM	0.00	0.00	0.00
Post-nonlinear causal model	18.17	28.83	51.63
Number of latent confounders $Q = 6$:			
Our approach (t-distributed individual-specific effects)	42.66	76.29	132.43
Our approach (Gaussian individual-specific effects)	33.13	69.07	104.83
LvLiNGAM (1 latent confounder)	2.40	2.56	9.38
LvLiNGAM (4 latent confounders)	22.17	30.12	83.01
SLIM (1 latent confounder)	5.89	6.22	6.77
SLIM (4 latent confounders)	7.58	8.18	9.11
SLIM (10 latent confounders)	11.03	12.02	13.91
LiNGAM-GC-UK	0.00	0.00	0.00
ICA-LiNGAM	0.06	0.05	0.05
DirectLiNGAM	0.01	0.01	0.01
Pairwise LiNGAM	0.00	0.00	0.00
Post-nonlinear causal model	18.71	29.62	52.21
Number of latent confounders $Q = 12$:			
Our approach (t-distributed individual-specific effects)	29.16	59.30	134.89
Our approach (Gaussian individual-specific effects)	32.18	68.14	104.76
LvLiNGAM (1 latent confounder)	2.35	2.50	13.58
LvLiNGAM (4 latent confounders)	21.51	30.10	94.08
SLIM (1 latent confounder)	5.90	6.03	6.62
SLIM (4 latent confounders)	7.58	7.99	8.97
SLIM (10 latent confounders)	10.92	11.68	13.74
LiNGAM-GC-UK	0.00	0.00	0.00
ICA-LiNGAM	0.07	0.08	0.07

DirectLiNGAM	0.01	0.02	0.02
Pairwise LiNGAM	0.00	0.00	0.00
Post-nonlinear causal model	18.21	29.21	51.89

Table 4: Average CPU time (s)

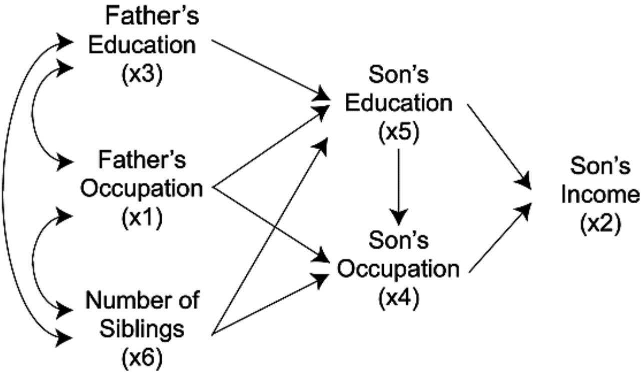


Figure 4: Status attainment model based on domain knowledge. Usually, the relations of x_1, x_3, and x_6, represented by bi-directed arcs, are not modeled.

causal directions among the three variables would be $x_1 \leftarrow x_3$, $x_6 \leftarrow x_1$, and $x_6 \leftarrow x_3$ based on their temporal orders.

Table 5 shows the numbers of successes and precisions. Our mixed-LiNGAM approach with the t-distributed individual-specific effects gave the largest number of successful discoveries 12 and achieved the highest precision, i.e., num. successes / num. pairs = $12/15 = 0.80$. The second best method was our mixed-LiNGAM approach with the Gaussian individualspecific effects, which found one less correct possible directions than the t-distribution version. The third best method was LvLiNGAM with 1 latent confounder, which found two less correct possible directions than the t-distribution version. This would be mainly because our two methods allow individual-specific effects and the other methods do not.

Table 6 shows the estimated hyper-parameter values of our mixed-LiNGAM approach with the t-distributed individual-specific effects that performed best in the sociology data experiment. Either the estimated hyper-parameter $\hat{\tau}_1^{indvdl}$ or $\hat{\tau}_2^{indvdl}$ that represents the magnitudes of individual differences was non-zero in all pairs except (x_4, x_5). The nonignorable influence of latent confounders was implied between the pairs (x_2, x_4), (x_2, x_6) and (x_3, x_6) since both $\hat{\tau}_1^{indvdl}$ or $\hat{\tau}_2^{indvdl}$ were non-zero for the pairs. In addition, for the pair (x_2, x_6), there might exist some nonlinear influence of latent confounders, since $\hat{\sigma}_{12}$ is zero, i.e. , the individual-specific effects were linearly uncorrelated but dependent.[16] If $\hat{\sigma}_{12}$ were larger, it would have implied a larger linear influence of the latent confounders on the pair (x_2, x_6). The estimates of the hyper-parameter τ_1^{indvdl} were very large for the pairs (x_2, x_6) and (x_4, x_1), which implied very large individual differences regarding x_2 and x_4 respectively. This might imply that the estimated directions could be less reliable, although they were correct in this example.

Another point is that both our methods with t-distributed and Gaussian individual-specific effects failed to find the possible direction $x_5 \leftarrow x_1$, although the causal relation is

16. Two variables that follow the multivariate t-distribution are dependent, even when they are uncorrelated, as stated in Section 3.2.

expected to occur from the domain knowledge (Duncan et al., 1972). This failure would be attributed to the model misspecification since the sample size was very large. Since the estimate of the hyper-parameter τ_1^{indvdl} regarding x_5 was zero, the influence of latent confounders might be small for this pair, although the estimate of τ_2^{indvdl} was not small and the individual difference regarding x_5 seemed substantial. Modeling both latent confounders and nonlinear relations and/or allowing a wider class of non-Gaussian distributions might lead to better performance. This is an important line of future research.

Possible directions | Our approach
LvLiNGAM SLIM
| t-dist. Gaussian Num.
lat. conf. Num. lat. conf.
| 1
4 1 4 10 $x_1(FO) \leftarrow x_3(FE)$ | ✓
✓ ✓ ✓ $x_2(SI) \leftarrow x_1$
(FO) | ✓ ✓
✓ ✓ $x_2(SI) \leftarrow x_3(FE)$ | ✓ ✓
✓ ✓ ✓ $x_2(SI) \leftarrow x_4(SO)$ | ✓
✓ ✓ ✓ $x_2(SI) \leftarrow x_5(SE)$ |
✓ ✓ ✓ ✓ ✓
✓ $x_2(SI) \leftarrow x_6(NS)$ | ✓ ✓ ✓
✓
$x_4(SO) \leftarrow x_1(FO)$ | ✓ ✓ ✓
✓ ✓ ✓ ✓ $x_4(SO) \leftarrow x_3(FE)$ | ✓
✓ ✓ ✓ ✓ ✓ $x_4(SO) \leftarrow x_5$
(SE) | ✓ ✓ ✓ ✓
$x_4(SO) \leftarrow x_6(NS)$ | ✓ ✓ ✓
✓ ✓ $x_5(SE) \leftarrow x_1(FO)$ |
✓ $x_5(SE) \leftarrow x_3(FE)$ | ✓ ✓
✓ ✓ ✓ $x_5(SE) \leftarrow x_6(NS)$ | ✓
✓ ✓ ✓ ✓ $x_6(NS) \leftarrow x_1(FO)$ |
✓ ✓ $x_6(NS) \leftarrow x_3(FE)$ |
✓ ✓ ✓ ✓

Num. of successes | 12 11 10 9 9 7 8
Precisions | 0.80 0.73 0.67
0.60 0.60 0.47 0.53

Possible directions | LiNGAM-GC-UK ICA Direct Pairwise PNL
$x_1(FO) \leftarrow x_3(FE)$ | ✓ ✓
$x_2(SI) \leftarrow x_1(FO)$ | ✓ ✓

✓ $x_2(SI) \leftarrow x_3(FE)$		✓		
✓ $x_2(SI) \leftarrow x_4(SO)$		✓		✓
✓ $x_2(SI) \leftarrow x_5(SE)$		✓		
✓ $x_2(SI) \leftarrow x_6(NS)$		✓		
✓ $x_4(SO) \leftarrow x_1(FO)$				✓
✓				
$x_4(SO) \leftarrow x_3(FE)$			✓	
✓ $x_4(SO) \leftarrow x_5(SE)$		✓		
✓ $x_4(SO) \leftarrow x_6(NS)$		✓		
$x_5(SE) \leftarrow x_1(FO)$	✓		✓	
✓				
$x_5(SE) \leftarrow x_3(FE)$			✓	
✓ $x_5(SE) \leftarrow x_6(NS)$				
$x_6(NS) \leftarrow x_1(FO)$	✓		✓	
$x_6(NS) \leftarrow x_3(FE)$	✓		✓	
✓				
Num. of successes	3 8	9 2 9		
Precisions	0.20	0.53	0.60	

0.13 0.60

FO: Father's Occupation ICA: ICA-LiNGAM (Shimizu et al., 2006) FE: Father's Education
Direct: DirectLiNGAM (Shimizu et al., 2011) SI: Son's Income
Pairwise: Pairwise LiNGAM (Hyvärinen and Smith, 2013) SO: Son's Occupation PNL: Post-nonlinear causal model SE: Son's Education (Zhang and Hyvärinen, 2009)
NS: Number of Siblings

Table 5: Comparison of eight methods

Pairs analyzed	Possible directions	Estimated directions	$\hat{\tau}_1^{indvdl}$
$(x_1(FO), x_3(FE))$	$\leftarrow$	$\leftarrow$	$0.4^2\widehat{\mathrm{var}}(x_1)$
$(x_2(SI), x_1(FO))$	$\leftarrow$	$\leftarrow$	$0.8^2\widehat{\mathrm{var}}(x_2)$
$(x_2(SI), x_3(FE))$	$\leftarrow$	$\leftarrow$	$0.8^2\widehat{\mathrm{var}}(x_2)$
$(x_2(SI), x_4(SO))$	$\leftarrow$	$\leftarrow$	$0.2^2\widehat{\mathrm{var}}(x_2)$
$(x_2(SI), x_5(SE))$	$\leftarrow$	$\leftarrow$	0
$(x_2(SI), x_6(NS))$	$\leftarrow$	$\leftarrow$	$1.0^2\widehat{\mathrm{var}}(x_2)$
$(x_4(SO), x_1(FO))$	$\leftarrow$	$\leftarrow$	$1.0^2\widehat{\mathrm{var}}(x_4)$
$(x_4(SO), x_3(FE))$	$\leftarrow$	$\leftarrow$	0
$(x_4(SO), x_5(SE))$	$\leftarrow$	$\leftarrow$	0
$(x_4(SO), x_6(NS))$	$\leftarrow$	$\leftarrow$	$0.6^2\widehat{\mathrm{var}}(x_4)$
$(x_5(SE), x_1(FO))$	$\leftarrow$	$\rightarrow$	0
$(x_5(SE), x_3(FE))$	$\leftarrow$	$\leftarrow$	$0.6^2\widehat{\mathrm{var}}(x_5)$
$(x_5(SE), x_6(NS))$	$\leftarrow$	$\leftarrow$	$0.2^2\widehat{\mathrm{var}}(x_5)$
$(x_6(NS), x_1(FO))$	$\leftarrow$	$\rightarrow$	$0.2^2\widehat{\mathrm{var}}(x_6)$
$(x_6(NS), x_3(FE))$	$\leftarrow$	$\rightarrow$	$0.2^2\widehat{\mathrm{var}}(x_6)$

FO: Father's Occupation
FE: Father's Education
SI: Son's Income
SO: Son's Occupation
SE: Son's Education

NS: Number of Siblings

$\hat{\tau}_2^{indvdl}$	$\hat{\sigma}_{12}$
0	-0.7
0	0.3
0	-0.5
$0.4^2\widehat{\mathrm{var}}(x_4)$	-0.5
$0.4^2\widehat{\mathrm{var}}(x_5)$	0

$0.6^2\widehat{\text{var}}(x_6)$	0
0	0.9
$0.2^2\widehat{\text{var}}(x_3)$	-0.3
0	-0.3
0	-0.7
$0.8^2\widehat{\text{var}}(x_1)$	0.3
0	-0.5
0	-0.3
0	-0.9
$0.6^2\widehat{\text{var}}(x_3)$	0.5

τ_1^{indvdl} and τ_2^{indvdl} represent the variances of the individual-specific effects for the variable pairs in the left-most column.
σ_{12} represents the correlation parameter value of the individual-specific effects for the variable pairs in the left-most column.

Table 6: Estimated hyper-parameter values of our method with t-distributed individual-specific effects

6. Conclusions and Future Work

We proposed a new variant of LiNGAM that incorporated individual-specific effects in order to allow latent confounders. We further proposed an empirical Bayesian approach to estimate the possible causal direction of two observed variables based on the new model. In experiments on artificial data and real-world sociology data, the performance of our method was better than or at least comparable to that of existing methods.

For more than two variables, one approach would be to apply our method on every pair of the variables. Then, we can estimate a causal ordering of all the variables by integrating the estimation results. This approach is computationally much simper than trying all the possible causal orderings. Once a causal ordering of the variables is estimated, the remaining problem is to estimate regression coefficients or their posterior distributions. Then, one can see if there are direct causal connections between these variables. Although this could still be computationally challenging for large numbers of variables, the problem reduces to a significantly simpler one by identifying their causal orders. Thus, it is sensible to develop methods that can estimate causal direction of two variables allowing latent confounders.

A reviewer suggested that we can generalize our model to more than two variables. Instead of a two-equation system in Table 1 we could have any number of equations each with an individual-specific confounder variable, although this approach would be computationally challenging.

Future work will focus on extending the model to allow cyclic and nonlinear relations and a wider class of non-Gaussian distributions as well as evaluating our method on various real-world data. Another important direction is to investigate the degree to which the model selection is sensitive to the choice of prior distributions.

Acknowledgments

S.S. was supported by KAKENHI #24700275. We thank Aapo Hyvärinen, Ricardo Silva and three reviewers for their helpful comments.

References

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

P. Billingsley. *Probability and Measure.* Wiley-Interscience, 1986.

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

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

D. M. Chickering and J. Pearl. A clinician's tool for analyzing non-compliance. In *Proc. 13th National Conference on Artificial Intelligence (AAAI1996)*, pages 1269–1276, 1996.

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

E. Demidenko. *Mixed Models: Theory and applications.* Wiley-Interscience, 2004.

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

O. D. Duncan, D. L. Featherman, and B. Duncan. *Socioeconomic Background and Achievement.* Seminar Press, New York, 1972.

D. Entner and P. O. Hoyer. 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, 2011.

J. Eriksson and V. Koivunen. Identifiability and separability of linear ICA models revisited. In *Proc. Fourth International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, pages 23–27, 2003.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

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

P. O. Hoyer and A. Hyttinen. Bayesian discovery of linear acyclic causal models. In *Proc.*

25th Conference on Uncertainty in Artificial Intelligence (UAI2009), pages 240–248, 2009.

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

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

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696. 2009.

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

A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001a.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001b.

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

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

S. Kotz and S. Nadarajah. *Multivariate t-distributions and Their Applications*. Cambridge University Press, 2004.

I. G. G. Kreft and J. De Leeuw. *Introducing Multilevel Modeling*. Sage, 1998.

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

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

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proc. 11th Conference on Uncertainty in Artificial Intelligence*, pages 411–418, 1995.

A. Moneta, N. Chlaß, D. Entner, and P. Hoyer. Causal search in structural vector autoregressive models. In *Journal of Machine Learning Research: Workshop and Conference Proceedings, Causality in Time Series (Proc. NIPS2009 Mini-Symposium on Causality in Time Series)*, volume 12, pages 95–114, 2011.

A. Moneta, D. Entner, P.O. Hoyer, and A. Coad. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75 (5):705–730, 2013.

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

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450, 2011a.

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

J. D. Ramsey, R. Sanchez-Romero, and C. Glymour. Non-Gaussian methods and high-pass filters in the estimation of effective connections. *NeuroImage*, 84(1):986–1006, 2014.

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

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

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

S.M. Smith, K.L. Miller, G. Salimi-Khorshidi, M. Webster, C.F. Beckmann, T.E. Nichols, J.D. Ramsey, and M.W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011.

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

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

P. Spirtes, C. Glymour, and R. Scheines.

Causation, Prediction, and Search. Springer Verlag, 1993. (2nd ed. MIT Press 2000).

P. Spirtes, C. Glymour, R. Scheines, and R. Tillman. Automated search for causal relations: Theory and practice. In R. Dechter, H. Geffner, and J. Halpern, editors, *Heuristics, Probability, and Causality: A Tribute to Judea Pearl*, pages 467–506. College Publications, 2010.

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

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

A. von Eye and L. R. Bergman. Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology*, 15(3):553–580, 2003.

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

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