Available online at www.sciencedirect.com

ScienceDirect

International Journal of Approximate Reasoning
49 (2008) 362–378

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Estimation of causal effects using linear non-Gaussian causal models with hidden variables

Patrik O. Hoyer *, Shohei Shimizu, Antti J. Kerminen, Markus Palviainen

Helsinki Institute for Information Technology, Department of Computer Science, University of Helsinki, Helsinki, Finland
Learning and Inference Group, The Institute of Statistical Mathematics, Japan

Available online 29 February 2008

Abstract

The task of estimating causal effects from non-experimental data is notoriously difficult and unreliable. Nevertheless, precisely such estimates are commonly required in many fields including economics and social science, where controlled experiments are often impossible. Linear causal models (structural equation models), combined with an implicit normality (Gaussianity) assumption on the data, provide a widely used framework for this task.

We have recently described how *non-Gaussianity* in the data can be exploited for estimating causal effects. In this paper we show that, with non-Gaussian data, causal inference is possible even in the presence of hidden variables (unobserved confounders), even when the existence of such variables is unknown a priori. Thus, we provide a comprehensive and complete framework for the estimation of causal effects between the observed variables in the linear, non-Gaussian domain. Numerical simulations demonstrate the practical implementation of the proposed method, with full Matlab code available for all simulations.

Keywords: Causal discovery; Structural equation models; Independent component analysis; Non-Gaussianity; Latent variables

1. Introduction

The ultimate goal of much of the empirical sciences is the discovery of *causal relations*, as opposed to just correlations (associations), between variables. In other words, one seeks to understand how various measured quantities or phenomena *influence* each other, so as to be able to predict the consequences of actions.

In many cases, causal effects can be estimated using controlled randomized experiments. A typical case is the testing of new medicines: subjects are randomly divided into treatment and control groups so that the groups are statistically identical *but for the actual medicine given*. Differences in responses between the groups can then be uniquely attributed to the effect of the medicine.

Unfortunately, however, in many fields of science and in many studies it is not always possible to perform controlled experiments. Often it can be very costly, unethical, or even technically impossible to directly control

* Corresponding author.
E-mail address: Patrik.Hoyer@cs.helsinki.fi (P.O. Hoyer).

0888-613X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.02.006

the variables whose causal effects we wish to learn. In such cases one must rely on observational studies combined with prior information and reasonable assumptions to learn causal relationships.

Linear causal models, also known as Structural Equation Models (SEM), can be thought of as the simplest possible causal models for continuous-valued data, and they have been the target of much research over the past decades, see e.g. [3,15] and references therein. The bulk of this research has, however, made an either explicit or implicit assumption of Gaussian data (i.e. normally distributed data). Although the methods that have been developed work for any distributions, these methods have been limited in their estimation abilities by what can be accomplished for Gaussian data. Fortunately, in the real-world, many variables are inherently non-Gaussian, and in such a case one can obtain much stronger results than for the Gaussian case.

In this contribution, we begin by giving a tutorial-level introduction to our previous results showing how non-Gaussianity can be useful for the estimation of causal effects [14]. We then take this idea further and show how, in this framework, confounding hidden variables can be detected and taken into account in the estimation method.

This article is organized as follows: First, in Section 2 we describe the estimation problem when there are only two variables which have been measured. This allows us to explain the basic idea as simply as possible, before proceeding to the general case in Section 3. Section 4 provides details needed for a practical implementation of the method, while Section 5 describes simulations that verify the theory and the algorithms developed. Finally, Sections 6 and 7 provide some discussion and give conclusions.

2. Two observed variables

2.1. Cause vs correlation

The prototypical problem of inferring causal effect from correlation (association) can best be illustrated in the simple case of only two measured variables. A classic example is the debate concerning the effect of playing violent video games. Based on studies showing a correlation between playing such games and violent behavior, many argue that video games encourage violence (see [1]). Others, however, point out that the correlation could equally well be simply explained by violent individuals enjoying such games. The crucial policy question, of course, is whether banning the games would reduce crime in society. Unfortunately, the answer depends on which of the two explanations fits actual human behavior better.

Another example concerns the statistical dependency between poverty and illiteracy. It is well documented that, over the set of nations, there is a strong correlation between low income (measured using Gross Domestic Product per person) and illiteracy in the adult population [18]. Yet this does not directly tell us which is the cause and which the effect. Additionally, it is possible, even likely, that there exists some background factor which accounts for both poverty and illiteracy. It is also plausible that the variables are mutually reinforcing. The reason we care about the causal relationships at work is of course that they are crucial when deciding on appropriate actions.

We will formalize the two-variable causal inference problem as follows. We observe S independent, identically distributed samples $\mathbf{x}^{(s)},\ s = 1,\ldots,S$, of a random variable $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ with a two-dimensional probability density $p(\mathbf{x}) = p(x_1, x_2)$ where x_1 and x_2 are the two (continuous-valued) quantities in question (for instance, $x_1 =$ GDP per person and $x_2 =$ percentage of adult population who are literate; S is the number of countries for which data is available). One of the following mechanisms is generating the data:

I x_1 and x_2 are not causally related and do not share any common cause.
II x_1 has an effect on x_2, and they have no common cause.
III x_2 has an effect on x_1, and they have no common cause.
IV x_1 and x_2 are statistically dependent solely due to a common cause.
V x_1 has an effect on x_2, and in addition they have a common cause.
VI x_2 has an effect on x_1, and in addition they have a common cause.

These possibilities have been illustrated as graphs in Fig. 1a, in which each variable is denoted by a node and directed edges (arrows) between nodes denote direct causal relationships. (At this point, please disregard the labels for the edges.)

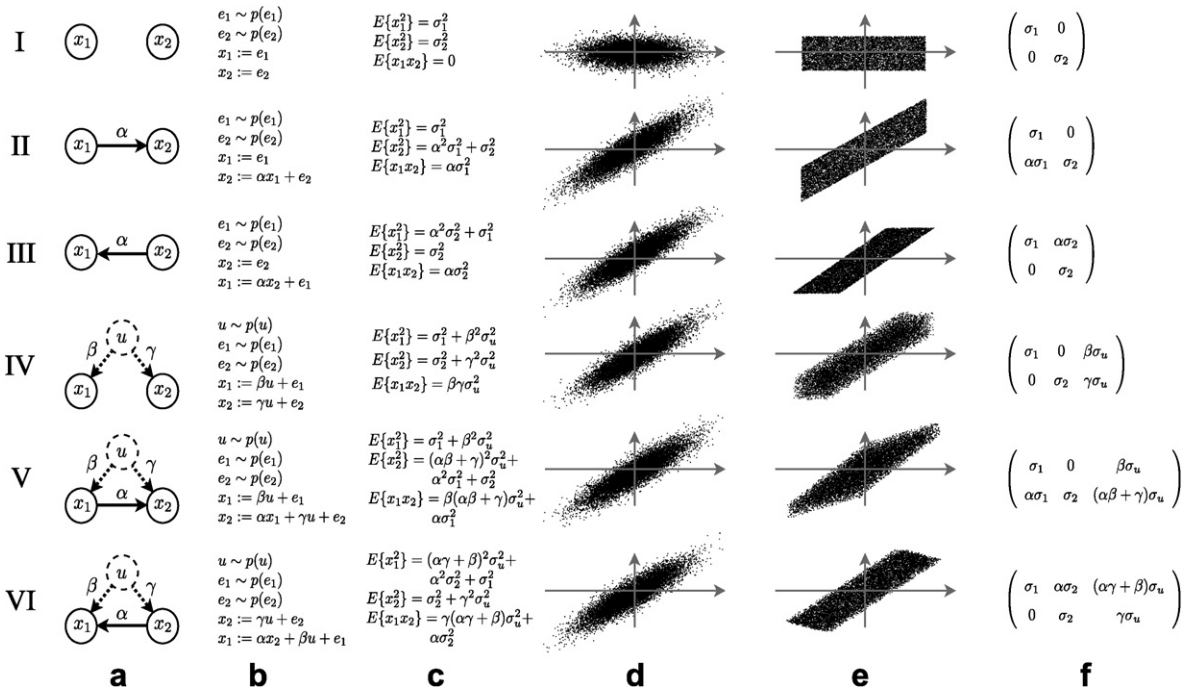

Fig. 1. (a) Models I–VI depicted as graphs. The observed variables (x_1 and x_2) and the common cause make up the nodes of the graphs, whereas the arrows denote direct causal relationships. The labels on the arrows denote the strength of the connections (explained in Section 2.2). (b) The explicit data generating process for each model. First, the disturbance variables (and the hidden common cause) are sampled *mutually independently* from their respective distributions, then the values of the observed variables are assigned in the causal order. (c) Variances and covariance produced by the different models. Note that each of the models II–VI can produce any valid covariance matrix by suitable choices of the parameters. (d) When the disturbances and common cause have Gaussian distributions, the full distribution is jointly Gaussian. Here we depict, using scatterplots, such distributions $p(x_1, x_2)$. For models II–VI we have specifically chosen parameters which yield the exact same covariance matrices and hence the exact same distributions. (e) The distributions $p(x_1, x_2)$ produced when the disturbances and the common cause have uniform distributions. The networks have the same parameter values as in (d). (f) The ICA matrices $\mathbf{A}$ corresponding to the different models. See main text (Sections 2.2 and 2.3) for details.

Several comments are in order. First, note that we are explicitly assuming that our data is an unbiased sample from the distribution $p(\mathbf{x})$. In other words, 'selection' effects [5] whereby the values of the variables may influence whether the data point is included in (or excluded from) the data are not considered in our framework.

A second important caveat is that we have ruled out feedback effects from our models. In other words, x_1 and x_2 cannot mutually reinforce each other. This might seem a major drawback with regards to applicability, as feedback might be considered quite plausible in many cases (including the examples presented above). This problem is considered further in the discussion in Section 6, but for now assume that no feedback is present.

Given the data set $\{\mathbf{x}^{(s)}\}$, the causal inference problem is to judge the plausibility of the various alternative models I–VI having generated the data. In the theory developed in this paper, it will be assumed that the number S of data vectors is large, in principle approaching infinity (so that we can neglect small sample effects). Nevertheless, the simulations will show that estimation is possible even for finite, manageable data sets.

2.2. Linear Gaussian causal models

In linear causal models (also known as Structural Equation Models), the relationships between the variables are described by linear functions. The value assigned to each variable is a weighted sum of its direct causes (parents in the graph) plus a 'disturbance' or 'error' term which makes the system non-deterministic. The linearity implies that associated with each directed edge is a (real-valued) coefficient that gives the weight of the parent to the value of the child. On the other hand, each observed node j has an associated (not directly

observed) disturbance variable e_j with some probability density $p(e_j)$, and the common cause u a distribution $p(u)$, and all the disturbances and the common cause are mutually independent. In Fig. 1a, the edges in the graphs are labeled by their corresponding coefficients (we do not explicitly show the disturbance variables or the probability densities), and in Fig. 1b we state the data generating process explicitly.

Before proceeding to the problem of inferring a model from the observed data, let us make sure that we are absolutely clear about the causal implications of the models. Pearl [13] has compellingly argued for the need for a distinction between making predictions based on *observations*, on the one hand, and predictions involving *interventions*, on the other. Our uncertainty regarding the value of x_2 given that we observed the value of x_1 is given by $p(x_2|x_1)$, the standard conditional probability distribution. However, if we were to intervene and set x_1 to some particular value, our uncertainty regarding the value of x_2 would be represented by $p(x_2|\mathrm{do}(x_1))$, which would be calculated from the model by first removing all edges pointing *into* x_1 and then proceeding as in the observational case [13].

In the linear models considered in this paper, the coefficients associated with the links define the causal effects in the system. For instance, in model V of Fig. 1, α represents the change in expected value of x_2 given a unit change in x_1 *in a situation where we are controlling the value of* x_1. Formally, $\alpha = \frac{\mathrm{d}}{\mathrm{d}x_1} E\{x_2|\mathrm{do}(x_1)\}$. This is to be clearly distinguished from the change in expected value of x_2 given a unit change in the *observed* value of x_1; this is given by $\frac{\mathrm{d}}{\mathrm{d}x_1} E\{x_2|x_1\}$ and only depends on the joint distribution $p(x_1, x_2)$ and not on the particular causal mechanisms generating that distribution.

As stated in the introduction, most work on linear causal models has made the (explicit or implicit) assumption that the data is Gaussian (follows the multivariate normal distribution). This assumption is convenient for many reasons. It implies that all information is contained in the mean and the covariance matrix, and no further statistics need to be obtained from the data. It also implies that all conditional and marginal distributions are Gaussian, as are any linear transformations. Further, statistical tests (for conditional uncorrelatedness, for instance) are well developed for the normal distribution. Finally, the central limit theorem is often used to argue that many real-world variables are likely to have Gaussian distributions.

The variances and covariance implied by the models I–VI have been calculated in Fig. 1c. (From here on, we assume for simplicity that all variables have zero-mean. This implies no loss of generality, as all data can be trivially put in this form by subtracting out the mean from each variable.) For Gaussian data, the causal inference problem is restricted by the degree to which these covariance matrices are distinguishable from each other. The essential question is, given that the data was generated with one of the models, which other models are equally compatible with the data? If the data exhibits a clear correlation between x_1 and x_2, it is clear that model I does not fit since it forces the two to be independent. However, all other models (II–VI) are compatible with a correlation between the variables. In fact, each of these models (II–VI) can generate *any* valid covariance matrix, for suitable choices of the parameters. Thus none of them could ever be completely ruled out. To illustrate this, we have plotted, in Fig. 1d, the Gaussian distributions yielded by the models. In particular, for models II–VI we have selected the values of the parameters so that the models produce identical variances and covariance.

Note that although in principle models II–VI cannot ever be ruled out, if the data actually exhibits zero correlation between x_1 and x_2, then these models could be rejected on the basis of *faithfulness* [17] (also called *stability* [13]): If the parameter values were selected randomly then a zero correlation is a very special accidental event. Thus, zero correlation in the data would uniquely imply model I as it is the only model whose *structure* implies uncorrelatedness. However, it must be kept in mind that faithfulness is just an assumption; it may well be violated in some systems, but it seems reasonable to assume it in most cases. It is, in effect, a version of Occam's razor, a well established general principle in science.

Nevertheless, even assuming faithfulness, it is clear that in this case of just two observed variables not much can be done in terms of causal inference for linear Gaussian models. The direction of causation cannot be determined, and neither can the presence of an unobserved common cause. Fortunately, we will see in the next subsection that when the models are non-Gaussian the situation is much better.

2.3. Exploiting non-Gaussianity

A non-Gaussian probability distribution is a distribution whose density takes any shape other than the classic bell-shaped normal curve. Such distributions are in general analytically much more difficult to work with than

Gaussians, particularly in the multi-dimensional case, since there are not usually any guarantees that linear transformations, marginalization, or conditionalization will result in distributions of the same parametric family.

Non-Gaussian distributions, however, exhibit much more structure than does the Gaussian. This structure can be useful for the estimation of causal effects. To illustrate the benefit non-Gaussianity yields, in Fig. 1e we give scatterplots of data produced when both of the disturbance variables e_j and the hidden common cause u all have zero-mean uniform distributions. The distributions produced by the different models (I–VI) are now all mutually distinguishable!

It may seem that the distinguishability of the distributions produced by the different models is a special result given by the properties of the uniform distribution. Or, at the very least, it might seem that the distributions are likely to be distinguishable only if we, a priori, know the one-dimensional distribution(s) involved. In fact, the truth is quite the opposite. The Gaussian is the special case, the peculiar properties of which leads to indistinguishability. Any type of arbitrary non-Gaussian distributions (unknown to us a priori) allow us, in principle, to differentiate the models.

To see this, we explicitly write the observed variables in terms of the disturbances and the hidden variable. Since the full system is linear, and the concatenation of a linear function with another linear function is still linear, we can write the full system as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{A} \begin{pmatrix} e_1 \\ e_2 \\ u \end{pmatrix} \tag{1}$$

where $\mathbf{A}$ is the transformation matrix from the disturbances (and hidden variable) to the observed variables. (Essentially, $\mathbf{A}$ contains the total effects of each disturbance variable onto each observed variable.) In Fig. 1f, we have calculated these matrices for models I–VI. Note that the size of $\mathbf{A}$ is 2×2 for cases I–III as there is no common cause u, but for IV–VI it is 2×3. Further, notice that models I and IV have zeros in both rows, models II and V have a zero in the top row only, and models III and VI have a zero in the bottom row only.

The faithfulness assumption mentioned in Section 2.2 plays an important role in our subsequent analysis, because the analysis will depend on the locations of zeros in these matrices. For specific choices of the parameters (unfaithful models) it is possible to produce *additional* zeros in the transformation matrix $\mathbf{A}$, possibly leading to confusion in identifying the correct model. Hence, we have to require all models to be faithful.

Assuming faithfulness, the six models I–VI are distinguishable if we can estimate the matrix $\mathbf{A}$. Fortunately, the well developed theory of *Independent Component Analysis* (ICA) [4,10] applies to our problem. Since the disturbance variables and the hidden variable are all presumed mutually independent and non-Gaussian, and since the transformation to the observed space is linear, the model is a classic case of ICA. When there is no common cause, we have the 'easy' case where the number of sensors equals the number of sources [4], but when there is a common cause we have the more difficult case of 'overcomplete basis' ICA [11,7]. Nevertheless, it has been proven that, without any prior knowledge on the distributions involved, the matrix $\mathbf{A}$ (known as the 'basis matrix' or the 'mixing matrix' in the ICA community) is identifiable given enough data [7]. (See Section 4 for a discussion on the practical aspects of this estimation.)

There are two important caveats here. First, it is impossible to determine to which degree the scale of the data (for each component) should be attributed to the variance of the independent component i, on the one hand, or the scale of the corresponding column of $\mathbf{a}_i$ of the matrix $\mathbf{A}$, on the other. This is not much of a problem here since it is really a question of model definition only. We have defined (without loss of generality) that each disturbance variable has unit weight (i.e. there is no weight parameter for the influence of the disturbance on its corresponding observed variable) and instead has an unrestricted variance, whereas the common cause u can be assumed (again without loss of generality) to have unit variance. This eliminates the scale indeterminacy.

The second and more important problem is that we are not able to determine the order of the columns. This is natural since completely independent variables can have no unique well defined order. We are always able to determine the model I–VI uniquely since the model identification based on $\mathbf{A}$ does not depend on the ordering of the columns: We can distinguish between models I–III vs IV–VI based on the *number* of columns of $\mathbf{A}$, and assuming faithfulness we can distinguish the models inside these groups by the presence of zeros in both or only one of its rows. Unfortunately estimating the parameters (coefficients) does in some cases depend on cor-

rectly identifying the columns. Specifically, in cases V and VI, there will be two different solutions for the parameters, only one of the two corresponding to the actual generating model. For instance, for model V, the causal effect of x_1 on x_2 (i.e. the value of α) cannot be uniquely identified. The two possibilities will be either α (the correct value) and $(\alpha\beta + \gamma)/\beta$ (an incorrect value). Nevertheless it is possible to know that the true value is one of these two possibilities. This problem will be discussed in detail in the next section which describes the general case of an arbitrary number of observed variables.

We conclude this section by considering the possibility that there may be more than one hidden common cause. That is, there could be multiple independent u_i, each with its own set of weights to the observed variables x_i. The effect of this would naturally be to increase the number of columns of **A** (one additional column for each additional common cause). Regardless of its size, the theory of ICA guarantees that **A** is nevertheless identifiable from the observed data [7]. This implies that we can identify the model graph (IV–VI, but with several common causes) correctly. However, when the graph contains a direct effect between the observed variables (i.e. we have case V or VI) then there will be $N_h + 1$ different solutions for the parameters, where N_h is the number of common causes, each solution corresponding to one specific choice for which of the columns containing no zeros is designated to belong to a disturbance variable, the rest associated with the confounders.

3. The general case

In most situations where inference of causal effects from non-experimental data is needed, there are more than just two measured variables. This is fortunate, as additional information can facilitate (but never impair, at least if used correctly) the inference procedure.

In some cases we are particularly interested in the causal relationship between one hypothetical cause and one putative effect, and all other measured variables are treated as associated variables which aid the process of causal inference. In particular, a technique widely used in statistical practice is to measure a large number of 'background' variables which are then 'controlled for' by including them in the regression equation used to estimate causal effect. However, it has been shown that this practice may in fact turn consistent estimates into inconsistent ones in some cases [16]. Thus, more sophisticated methods are urgently needed.

In other situations the researcher does not have any particular pair of variables in mind but, rather, is interested in finding novel causal explanations over the whole set of variables. These types of cases might well be referred to as problems of 'causal discovery' [17]. As discussed in this paper, methods based on a Gaussian framework are well developed for this problem, but a non-Gaussian approach allows more structure to be uncovered.

3.1. Model definition

Assume that we observe data generated by a linear, non-Gaussian, acyclic causal model but that we only observe a subset of the involved variables. That is, the process has the following properties:

1. The full set of variables (including any unobserved variables) $x_i,\ i = \{1, \ldots, m\}$ can be arranged in a *causal order*, such that no later variable causes any earlier variable. We denote such a causal order by $k(i)$. That is, the generating process is *recursive* [3], meaning it can be represented by a *directed acyclic graph* (DAG) [13,17].
2. The value assigned to each variable x_i is a *linear function* of the values already assigned to the earlier variables, plus a 'disturbance' (error) term e_i, and plus an optional constant term c_i, that is
$$x_i := \sum_{k(j)<k(i)} \tilde{b}_{ij} x_j + e_i + c_i. \tag{2}$$
3. The disturbances e_i are all zero-mean continuous random variables with *non-Gaussian* distributions, and the e_i are independent of each other, i.e. $p(e_1, \ldots, e_m) = \prod_i p_i(e_i)$.
4. The *observed variables* is a subset of the x_i. We denote the set containing the indices of the observed variables by $J \subseteq \{1, \ldots, m\}$. In other words, our data set contains only the $x_j, j \in J$.

Fig. 2a shows an example of such a *latent variable LiNGAM* (lvLiNGAM) model. Please note that the unobserved set of the x_i here play the role of the hidden variable(s) u in Section 2.

As described in Section 2, we further require the model to be faithful [17]. This means that there is no exact canceling of effects. That is, when multiple causal paths exist from one variable to another their combined effect does not equal exactly zero.

Finally, we assume that we are able to observe a large number of data vectors $\mathbf{x}^{(s)}$ (which contain the observed variables $x_j, j \in J$), and that each is generated according to the above described process, with the same set of observed variables J, same causal order $k(i)$, same coefficients $\tilde{b}_{ij}$, same constants c_i, and the disturbances e_i sampled independently from the same distributions.

The key difference to our previous work [14] is that we now allow *unobserved confounding variables*: hidden variables which affect multiple observed variables and hence are potentially problematic for any causal analysis. The key difference to other existing research involving linear models with latent variables, such as [3,15], is our assumption of *non-Gaussian* variables, allowing us to utilize methods based on higher-order statistics.

3.2. Canonical models

Consider the example model shown in Fig. 2a. It should be immediately clear that it is impossible to estimate the full generating model from a sample of data vectors $\mathbf{x} = (x_1, x_2, x_3, x_5, x_8)^{\mathrm{T}}$. This can be seen most clearly by the fact that the data generating model contains a hidden variable (x_7) with no observed descendants; detecting the presence of such a variable from our data is obviously not feasible since it has absolutely no effect on the observed data. Fortunately, the impossibility of detecting x_7 and of estimating the strengths of any connections to it are not cause for concern. The reason for this is that the variable is not in any way relevant with respect to our goal: finding the causal relationships between the observed variables.

Another causally irrelevant hidden variable is x_6, which simply mediates the influence of x_2 and x_4 onto x_8. We have

$$x_6 := 2x_2 + 4x_4 + e_6$$
$$x_8 := 3x_6 + e_8$$

leading to

$$x_8 := 3(2x_2 + 4x_4 + e_6) + e_8 = 6x_2 + 12x_4 + 3e_6 + e_8 = 6x_2 + 12x_4 + e'_8$$

where we have simplified $e'_8 = 3e_6 + e_8$. The resulting model is shown in Fig. 2b. Since the variances of the disturbances (and the particular forms of the densities) are unconstrained, this shows that we can remove the hidden variable x_6 from the model to obtain a simpler model in which *the observed data is identical* to that given by the original model and, in addition, *all the causal implications are identical as well*. As an example, the causal effect of x_2 on x_8 is, both in the original model and in the simplified model, given by $\frac{\mathrm{d}}{\mathrm{d}x_2} E\{x_8|\mathrm{do}(x_2)\} = 6$. Similarly, all other expressions of the form $\frac{\mathrm{d}}{\mathrm{d}x_i} E\{x_j|\mathrm{do}(x_i)\}$ and $\frac{\mathrm{d}}{\mathrm{d}x_i} E\{x_j|x_i\}$ are identical for the two models. Indeed, not only the expectations but the full probability distributions (both with and without interventions) are

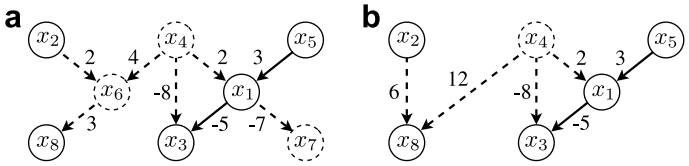

Fig. 2. (a) An example of a latent variable LiNGAM model. The diagram corresponds to the following data generation process: First the e_i are sampled mutually independently from their respective non-Gaussian distributions $p_i(e_i)$, after which the x_i are assigned through $x_2 := e_2$, $x_4 := e_4$, $x_5 := e_5$, $x_6 := 2x_2 + 4x_4 + e_6$, $x_1 := 2x_4 + 3x_5 + e_1$, $x_8 := 3x_6 + e_8$, $x_3 := -8x_4 - 5x_1 + e_3$, and $x_7 := -7x_1 + e_7$. (Here, all the c_i are zero, but this is not the case in general. Note that as in Fig. 1 we do not show the e_i or the probability densities in this graphical representation of the model.) Hidden variables are shown dashed; the observed data vector $\mathbf{x}$ consists of only the values of x_1, x_2, x_3, x_5, and x_8. (b) The canonical model corresponding to the network in (a). This is the observationally and causally equivalent network where all causally irrelevant variables have been simplified away. See main text for details.

in fact identical. Note that hidden variable x_4 cannot be similarly simplified away without either changing the observed data or allowing for dependencies between the disturbance variables, neither of which are acceptable in our framework.

We formalize this idea of causally relevant versus irrelevant hidden variables using the following concepts.

Two latent variable LiNGAM models are *observationally equivalent* if and only if the distribution $p(\mathbf{x})$ of the observed data vector is identical for the two models. This implies that the models cannot be distinguished based on observational data alone.

Two latent variable LiNGAM models are *observationally and causally equivalent* if and only if they are observationally equivalent and all causal effects of observed variables onto other observed variables are identical for the two models. In the notation of Ref. [13] these causal effects are given by $p(\mathbf{x}_{J_1}|\mathrm{do}(\mathbf{x}_{J_2})), J_1, J_2 \subseteq J$, with $J_1 \cap J_2 = \emptyset$. When these are identical for all choices of J_1 and J_2 the two models in question cannot be distinguished based on any observations nor any controlled experiments.

Finally, we define a *canonical model* to be any latent variable LiNGAM model where each latent variable is a root node (i.e. has no parents) and has at least two children (direct descendants). Furthermore, although different latent variables may have the same sets of children, no two latent variables exhibit exactly proportional sets of connection strengths to the observed variables. Finally, each latent variable is restricted to have zero-mean and unit variance.

To derive a canonical model which is observationally and causally equivalent to any given latent variable LiNGAM model, we can use the following algorithm:

Algorithm A. Given a latent variable LiNGAM model, returns an observationally and causally equivalent canonical model

1. First remove any latent variables without children. Iterate this rule until there are no more such nodes.
2. For any connection of the form $X \to Y$, where Y is a latent variable: (a) For all children Z of Y, add a direct connection $X \to Z$, the strength of the connection being the product of the strengths of $X \to Y$ and $Y \to Z$. If a direct connection already existed, add the specified amount to the original strength. (b) Remove the connection $X \to Y$. Iterate this rule until there are no more applicable connections.
3. Remove any latent variable with only a single child, incorporating the latent variable's disturbance variable and constant into those of the child. Iterate this until there are no more such latent variables.
4. For any pair of latent variables with *exactly proportional* sets of connection strengths to the observed variables, combine these into a single latent variable. Iterate this until there are no more such pairs of latent variables.
5. Finally, standardize the latent variables to have zero-mean and unit variance by adjusting the connection strengths to, and the constants of, their children.

The value of the concept of canonical models is that when searching for explanations of some given data we can, without loss of generality, restrict the search to such models. Any lvLiNGAM model can be fully represented by its observationally and causally equivalent canonical model, in the sense that *any* predictions (based on either observations or interventions) derived from the canonical model also hold for the original model. As we have mentioned in Section 2, and will discuss in more detail in Section 3.4, there will often be several causally distinct yet observationally equivalent canonical models. In such cases, to the extent that these canonical models differ from each other one cannot make definite causal predictions. Nevertheless, at no point is there any need to explicitly consider non-canonical models.

3.3. Model estimation in the Gaussian framework

As in Section 2, we start by briefly describing what can be achieved when the data follows the Gaussian distribution. In this case, causal inference must rely on conditional independencies in the distribution, as there is no additional structure in the data.

First, consider the case where the time order of the variables is known (essentially we know $k(i)$ as defined in Section 3.1) and it is known that there are no hidden variables. In this case there is a single model consistent with the order which produces the observed covariance matrix. Essentially, when the time order is known and there are no hidden variables, the causal estimation problem reduces to a statistical estimation problem. This is the basic idea behind the concept of Granger causality [8].

If it is known that there are no hidden variables, but there is no time ordering of the variables available, then there are often multiple distinct models that can explain a given covariance structure. This is due to *d-separation equivalence*: many different network structures may all imply the same set of conditional independencies; hence it is generally not possible to uniquely identify the generating model. The alternatives can be estimated using the PC-algorithm [17] (see also the IC algorithm of [13]).

In more realistic causal inference situations, it cannot usually be known that there are no hidden common causes. Thus, more advanced methods are required. One such method is the Fast Causal Inference (FCI) algorithm [17] which, given the set of conditional independencies valid in the data, outputs a set of causal relationships valid for all faithful models which could have generated the data. Without going into the details here we simply note that, because it is based on conditional independence properties alone, it is quite restricted in the amount of information it can output. As we saw in Section 2, making use of non-Gaussianity (where such exists) can yield a significant improvement in the ability to differentiate models. We now describe this approach in the general setting.

3.4. Model estimation based on non-Gaussianity

In this section we show how, from data generated by any faithful latent variable LiNGAM model, to estimate the set of canonical models which are observationally equivalent to the generating model.

We begin by considering the full data vector $\widetilde{\mathbf{x}} = \{x_1, \ldots, x_m\}$, which includes the latent variables. If we first subtract out the means of the variables, then the full data satisfies $\widetilde{\mathbf{x}} = \widetilde{\mathbf{B}}\widetilde{\mathbf{x}} + \mathbf{e}$, where, because of the DAG assumption, $\widetilde{\mathbf{B}}$ is a matrix (containing the *direct* effects between the full set of variables) that could be permuted to strict lower triangularity if one knew a causal ordering $k(i)$ of the variables. Solving for $\widetilde{\mathbf{x}}$ one obtains $\widetilde{\mathbf{x}} = \widetilde{\mathbf{A}}\mathbf{e}$, where $\widetilde{\mathbf{A}} = (\mathbf{I} - \widetilde{\mathbf{B}})^{-1}$ contains the influence of the disturbance variables onto the observed variables (the *total* effects). Again, $\widetilde{\mathbf{A}}$ could be permuted to lower triangularity (although not *strict* lower triangularity) with an appropriate permutation $k(i)$. Taken together, the linear relationship between $\mathbf{e}$ and $\widetilde{\mathbf{x}}$ and the independence and non-Gaussianity of the components of $\mathbf{e}$ define the standard linear *Independent Component Analysis* model [4,10].

So far, this is just restating what we pointed out in our previous work [14]. Now consider the effect of hiding some of the variables. This yields $\mathbf{x} = \mathbf{A}\mathbf{e}$, where $\mathbf{A}$ contains just the rows of $\widetilde{\mathbf{A}}$ corresponding to the observed variables. When the number of observed variables is less than the number of disturbance variables, $\mathbf{A}$ is non-square with more columns than rows. This is known as an *overcomplete* basis in the ICA literature [11,7].

Let us take a closer look at the structure inherent in the 'mixing matrix' $\mathbf{A}$. First, note that since for every latent variable LiNGAM model there is an observationally and causally equivalent canonical model, we can without loss of generality restrict our analysis to canonical models. Next, arrange the full set of variables such that all latent variables come first (in any internal order) followed by all observed variables (in a causal order), and look at the structure of the full matrix $\widetilde{\mathbf{A}}$ shown in Fig. 3a. Although we only observe part of the full

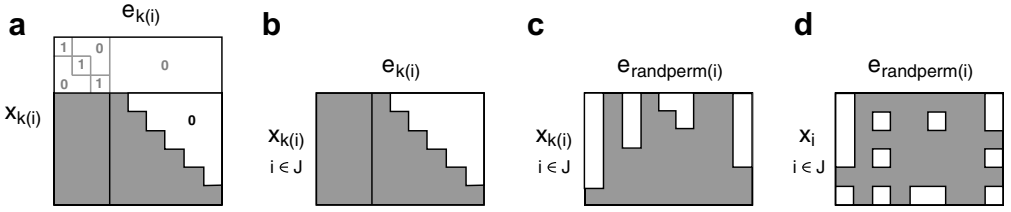

Fig. 3. (a) Basic structure of the full ICA matrix $\widetilde{\mathbf{A}}$ for a canonical model. (In this example, there are 3 hidden and 6 observed variables.) The top rows correspond to the hidden variables. Note that the assumption of a canonical model implies that this part of the matrix consists of an identity submatrix and zeros. The bottom rows correspond to the observed variables. Here, the shaded area represents values which, depending on the network, *may be zero or non-zero*. The white area represents entries which are zero by the DAG-assumption. (b) Since, by definition, we do not observe the latent variables, the information that we have is limited to the bottom part of the matrix shown in (a). (c) Due to the permutation indeterminacy in ICA, the observed basis matrix has its columns in random order. (d) Since we do not (up front) know a causal order for the observed variables, the observed mixing matrix $\mathbf{A}$ has the rows in the order of data input, not in a causal order.

matrix, with randomly permuted columns and with arbitrarily permuted rows as in Fig. 3d, the crucial point is that the observed ICA basis matrix **A** contains all of the information contained in the full matrix $\widetilde{\mathbf{A}}$ in the sense that all the free parameters in $\widetilde{\mathbf{A}}$ are also contained in **A**. Thus there is hope that the causal model could be reconstructed from the observed basis matrix **A**.

Of course we must remember that we do not directly observe **A** but must infer it from the sample vectors. Eriksson and Koivunen [7] have recently shown that the overcomplete ICA model is identifiable given enough data, and several algorithms are available for this task, see e.g. [12,2,10], extending the early results of Ref. [4] for the standard (square) mixing matrix ICA model. Thus, the remainder of this section considers the inference of the causal model were the exact ICA mixing matrix **A** known. The next section deals with the practical aspect of how to handle the inevitable estimation errors.

We again emphasize that ICA uses *non-Gaussianity* (that is, more than covariance information) to estimate the mixing matrix **A**. For Gaussian disturbance variables e_i, ICA cannot in general find the correct mixing matrix because many different mixing matrices yield the same covariance matrix, which in turn implies the exact same Gaussian joint density. Our requirement for non-Gaussianity of disturbance variables stems from the same requirement in ICA.

As in the standard square ICA case, identification in the overcomplete case is only up to permutation and scaling of the columns of **A**. The scaling indeterminacy is not serious; it simply amounts to a problem of not being able to attribute the magnitude of the influence of a disturbance variable e_i to its variance, the strength of its connection to its corresponding variable x_i, and in the case of hidden variables the average strength of the connections from that hidden variable to the observed variables. This is of no consequence since we are anyway never able to directly monitor the hidden variables nor the disturbance variables, making the scaling simply a matter of definition, as discussed in Section 2.3.

In contrast, the permutation indeterminacy *is* a serious problem, and in general leads to non-uniqueness in inferring the model: We cannot know which columns of **A** correspond to the hidden variables. Note, however, that this is the only information missing, as illustrated in Fig. 3. Thus, an upper bound for the number of observationally equivalent canonical models is the number of classifications into observed vs hidden. This is simply $(N_o + N_h)!/(N_o!N_h!)$, where N_o and N_h denote the numbers of observed and hidden variables.

Thus, we can formulate the following simple algorithm for calculating all observationally equivalent canonical models compatible with any given ICA basis matrix **A** (containing exact zeros):

Algorithm B. Given an overcomplete basis **A** (containing exact zeros) and the means of the observed variables, calculates all observationally equivalent canonical latent variable LiNGAM models compatible with the basis

1. N_h is determined as the number of columns of **A** minus the number of rows.
2. For each possible classification of the columns of **A** as belonging to disturbance variables of observed vs hidden variables:
   (a) Reorder the columns such that the ones selected as 'hidden variables' come first.
   (b) Augment the basis by adding the unobserved top part of the matrix (as in Fig. 3a), obtaining an estimate of $\widetilde{\mathbf{A}}$.
   (c) Test whether it is possible to permute $\widetilde{\mathbf{A}}$ (using independent row and column permutations) to lower triangular form. If not, go to the next classification.
   (d) Divide each column of $\widetilde{\mathbf{A}}$ by its diagonal element, and calculate the connection strengths $\widetilde{\mathbf{B}} = \mathbf{I} - \widetilde{\mathbf{A}}^{-1}$.
   (e) Check that the found network is compatible with the *faithfulness* assumption. If not, go to the next classification.
   (f) Add the found network $\widetilde{\mathbf{B}}$ to the list of observationally equivalent models which could have generated the data.

The practical problem of applying the above algorithm is that in actual data-analysis we do not know the matrix **A** exactly, but rather must estimate it from data. What ICA algorithms are most useful here, and how are we to deal with the inevitable estimation errors? These are questions we turn to next.

4. Practical aspects involving ICA estimation

Section 3 considered the idealized theoretical problem of calculating the possible models given the exact ICA basis matrix **A**. In this section we consider the practical problems which arise when the basis matrix has to be estimated from sample data. That is, we consider the spectrum of ICA algorithms available and

the problem of correctly identifying the zeros of the basis matrix from an estimate of it. Furthermore, since the structure of the network is related to the *inverse* of **A**, small estimation errors will give rise to small direct effects where none exist in the generating model. We describe our solution to these issues in this section. The full Matlab code implementing our solution is available on the internet, see the next section for details.

Let us first consider the choice of algorithm for estimating overcomplete ICA bases from data. For the purposes of our application, we seek accuracy, adaptability (as we in general do not know the densities of the disturbance variables in advance), and the possibility to estimate the model order (the number of columns in **A**). High-dimensionalities are out of the question so good scalability is probably not needed. Based on these considerations, the mixture of Gaussians framework [12,2] seems the most promising of the currently available alternatives [10]. In this method, the distributions of the disturbance variables are modeled as adaptable mixtures of Gaussians, allowing maximum likelihood estimation of all model parameters with no approximations.

It is important to note that the accuracy of the ICA estimate will depend on a number of factors. First, if the distributions of the disturbance variables are very close to Gaussian, it will in practice be impossible to obtain accurate estimates without a prohibitive number of data points. On the other hand, even if the distributions involved are significantly non-Gaussian, if they cannot be reasonably fit with the Gaussian mixture model, the resulting accuracy may not be very good. Another type of problem arises if the disturbance variables have very different variances. Such large differences of scale will be reflected in the norms of the columns of **A**, and this can lead to difficulties in estimation. Finally, because our method will depend on correctly identifying the zeros of **A** (see below), models which are close to unfaithful might be problematic.

Estimating the model order (i.e. the number of hidden variables) is a classic problem of model selection, and we do not offer any novel solution in this regard. In our initial implementation we have simply chosen to leave part of the data out of the learning phase and use that part for measuring the fit (likelihood) of the various models. This worked reasonably well but did not completely prevent overfitting.

Since our ICA algorithm does not yield exact zeros we have to be able to *infer* which coefficients in **A** are zero. To solve this problem we need to have some information on the uncertainty inherent in our estimates. We have employed the statistical method of bootstrapping [6], obtaining a *set* of estimates $\mathbf{A}_i$ representing our uncertainty regarding the elements of the mixing matrix. (Here, it is important to note that, because the columns are in an arbitrary permutation in each estimate, we must take care to match them with each other to obtain a proper estimate of the variance in **A**.) Once we have a matched set of basis matrices, we calculate the mean μ_{ij} and standard deviation σ_{ij} of each element A_{ij}, and heuristically set the probability of being zero for that element to

$$P(A_{ij} = 0) = \exp\left(-\frac{(\mu_{ij}/\sigma_{ij})^2}{2\alpha^2}\right), \tag{3}$$

where α is an adjustable parameter (we set $\alpha = 4.3$ based on some preliminary experiments). Then we go through all possible choices for zeros vs non-zeros, in order of decreasing probability, until we find a choice which yields a valid model. That is, we start by setting to zero all those elements A_{ij} for which $P(A_{ij} = 0) > 0.5$, and apply Algorithm B. If this does not produce a valid model, we flip the zero/non-zero status of the element whose probability is closest to 0.5, and try again. Assuming independence between the elements, we step through all possible classifications (there are 2^{NM} where N is the number of columns and M is the number of rows of **A**) in order of decreasing joint probability until we find a classification producing a valid model. Typically the inference of zeros worked well enough that the first choice (maximum probability) already yielded a valid model.

Finally, we want to prune insignificant direct effects in the resulting model. Note that we have until now only considered identifying the zeros of the matrix **A**, which represents the total effects of the disturbance variables on the observed variables. However, the absence of direct connections between the observed nodes are actually represented by zeros in $\widetilde{\mathbf{B}}$. To identify such zeros, we calculate $\widetilde{\mathbf{B}}_i$ corresponding to each bootstrapped $\mathbf{A}_i$ separately, and from the set of these we calculate the mean and the standard deviation (over the bootstrap repetitions) of each element of $\widetilde{\mathbf{B}}$. We then prune (set to zero) any elements of $\widetilde{\mathbf{B}}$ for which the absolute value of the mean is smaller than the standard deviation. This removes most of the insignificant direct connections between the nodes in the network.

In summary, although there are significant practical problems in estimating the overcomplete ICA basis matrix, identifying the zeros, and inferring the model, the problems can be overcome by the methods described above. The next section shows empirically that the estimation of small models from reasonably sized datasets, although difficult, is feasible.

5. Simulations

We have performed extensive simulations in order to (i) verify the algorithms described, and (ii) demonstrate in practice the estimation of small latent variable LiNGAM model from data. Full well documented Matlab code for all of these experiments is available at

`http://www.cs.helsinki.fi/group/neuroinf/lingam/lvlingam.tar.gz`

to allow the reader to effortlessly reproduce and verify our results.

First, we tested the theoretical concepts. We generated moderately sized (2–20 variables, 0–4 of which were hidden) random latent variable LiNGAM models. For each model, we analytically calculated its corresponding ICA basis matrix **A**. We then inferred, from each basis matrix, the set of observationally equivalent canonical models using Algorithm B. Depending on the network structure, the set consisted of either a single model or a small (2–6) number of observationally equivalent models. In every case, exactly one model in the inferred set was causally equivalent to the original model. Fig. 4 gives a typical example: The original (non-canonical) model is shown in (a), whereas the two canonical models which are observationally equivalent to this original model are given in (b) and (c). Of these, only (b) is also causally equivalent to the original model.

Second, and more importantly, we tested the practical ability of our system to estimate small networks from data. Because of the computational requirements of the overcomplete ICA estimation algorithm used [2], the extensive use of bootstrapping in estimation, and the goal of obtaining a large number of trials, only very small models (two or three observed variables, zero or one hidden confounders) were used. For the disturbance variables, we used four different distributions: The Laplace distribution, the χ^2-distribution with 3 degrees of freedom, and two different two-component Gaussian mixtures, one of which was supergaussian with a high peak and heavy tails, whereas the other was subgaussian and closer to a uniform distribution. See Fig. 5 for plots of the densities of these distributions. For each model, each disturbance variable was

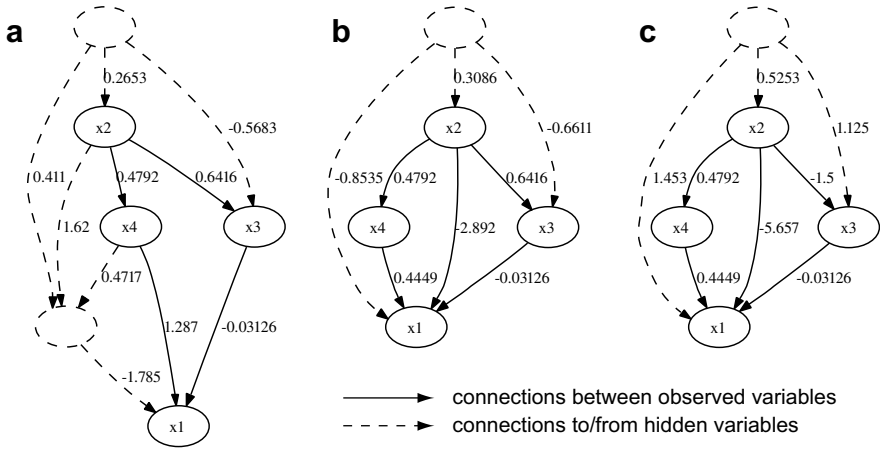

Fig. 4. (a) A randomly generated latent variable LiNGAM model. This model is not canonical, as there is a hidden variable which has parents in the graph. (b) The canonical model which is causally and observationally equivalent to the network in (a). Note that the connection from x_4 to x_1 is of strength $0.4449 \approx 1.287 + 0.4717 * (-1.785)$ (inaccuracies are due to rounding). Similarly, the direct connection from x_2 to x_1 is $-2.892 \approx 1.62 * (-1.785)$. (c) An observationally equivalent (but causally *not* equivalent) model. From the ICA basis which model (a) generates, it is impossible to distinguish between models (b) and (c). Hence we cannot determine all parameters in this case. Nevertheless, these are the only two alternatives, and note that many of the connection strengths are equal in the two models and thus uniquely determined.

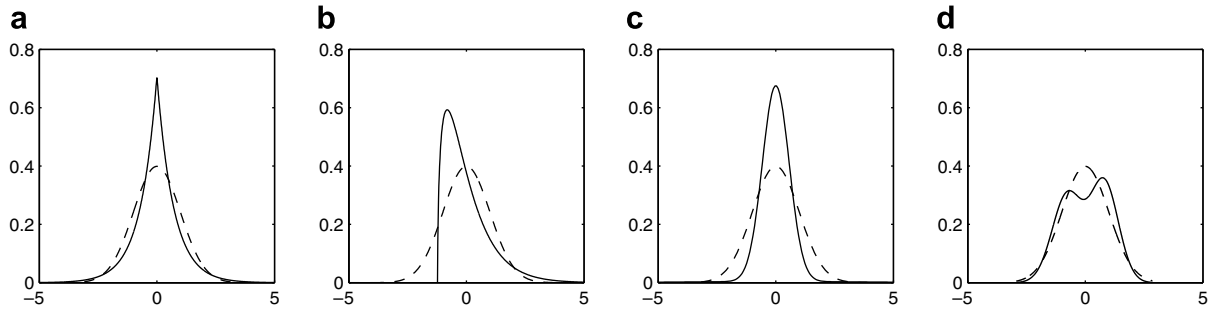

Fig. 5. Densities of disturbance variables (solid line) used in the simulations, compared with the density of the standard normal distribution (dashed line). (a) The Laplace distribution, $p(e_i) = \exp(-\sqrt{2}|e_i|)/\sqrt{2}$. (b) The χ^2-distribution with 3 degrees of freedom, i.e. $p(e_i) = \sqrt{e_i}\exp(-e_i/2)/\sqrt{2\pi}$, here shown standardized to zero-mean and unit variance. (c) A zero-mean, unit variance supergaussian two-component mixture of Gaussians density with a peak at zero and very heavy tails. The kurtosis of this distribution is approximately 40. (Note that the thickness of the tails is not seen very well in the plot, but can be inferred from the fact that the distribution has unit variance just like the standard normal.) (d) A zero-mean, unit variance subgaussian two-component mixture of Gaussians density. This distribution has a kurtosis close to −0.8 (i.e. it is more 'flat' than the normal distribution, as can be seen from the plot).

assigned one of these distributions, and a total of 10,000 data vectors were created by sampling the disturbance variables and generating the observed variables by the linear acyclic model.

For each of the generated datasets, we took a subset of 2500 data vectors and estimated both a square (3×3) ICA basis matrix (corresponding to a model with no hidden variable) and an overcomplete (3×4) basis matrix (one hidden variable) using maximum likelihood estimation with adaptable Gaussian mixtures [2]. The sources were modeled as two-component Gaussian mixtures; the true distributions of the disturbances were not known to the algorithm. (Note that the Laplace- and the χ^2-distributions can only be approximated by this model.) The likelihoods of both of the estimated models were evaluated against the remaining 7500 data vectors and the model with higher likelihood was selected. As described in Section 4, we subsequently bootstrapped the full dataset and estimated several basis matrices $\mathbf{A}_i$, used the variability in these estimates to identify zeros in the basis and thus infer the model, and finally pruned insignificant elements in the estimated coefficient matrix $\widetilde{\mathbf{B}}$. For each three-variable model, the whole estimation procedure, including all bootstrapping, took on the order of six to seven hours on a 2.8 GHz Pentium 4 processor-equipped PC.

In Fig. 6, we show typical results. The left column contains the original network used for generating the data, whereas the right column gives the set of models estimated from the datavectors without any knowledge of the true model. In cases (a), (b), (d) and (e), the disturbances were Laplace- and χ^2-distributed, whereas in case (c) the disturbances were two-component Gaussian mixtures (see Fig. 5). First, we note that, overall, the estimation works relatively well. In the five shown cases, all networks except case (d) were reasonably well estimated. Although (a) is the only network which results in a single model whose structure is identical to the original, we may also count (b), (c), and (e) as successes, as will be explained in detail below.

First, look at case (a). In the generating model (shown on the left), x_1 has a positive causal effect on x_2, and these two variables are causally unrelated to x_3. Nevertheless, a hidden confounder causes a correlation between x_2 and x_3. The estimated model (on the right) has the same structure and the connection strengths are very similar to those in the true model as well. Thus, causal predictions made on the basis of the estimated model would be quite accurate.

Next, consider model (b). In this case, there is a strong positive effect of x_2 on x_1, and again x_3 is causally independent in the true model (left column). Please note that in contrary to the previous case, now there is no hidden confounder and in fact x_3 is statistically independent of x_1 and x_2. The estimated model (right column) correctly represents the causal connection from x_2 to x_1, as well as the causal independence of x_3. There is, however, a hidden variable connecting x_1 with x_3 which did not exist in the generating model. Nevertheless this hidden confounder contributes only an insignificant amount of correlation because of the very weak connection to x_3. (Although this cannot be seen in the graph, the added variance in x_1 caused by the confounding variable is compensated by a correspondingly smaller variance of e_1.) Thus any observational or causal predictions derived from the estimated model are reasonably accurate.

The true model in panel (c) exhibits a negative causal effect of x_1 on x_2, and a positive causal effect of x_2 on x_3 (hence, the total effect of x_1 on x_3 is negative). A hidden variable confounds the relationship between x_1 and x_2. For this dataset, our method returned two possible models. This is because these two canonical models are observationally equivalent to each other, although they represent different causal relationships. First, examine the model on the right. Our model clearly has correctly found the causal relationships between the variables. Compared with the true model, the hidden confounder is incorrectly considered to directly influence x_3, but the strength of the effect is very small. It may also be noted that the effects of the confounding variable on x_1 and x_2 are of the wrong sign (compared with the generating model), but since we are never able to observe this variable its sign is arbitrary. (It should be noted though that the strengths of these connections are somewhat different than in the true model.) The alternative estimated model shows a positive effect of x_1 on x_2 (and hence also a positive effect on x_3), in clear violation of the true model. If we did not know the true model (as would always be the case in an application of the suggested method) we could not know which of the estimated

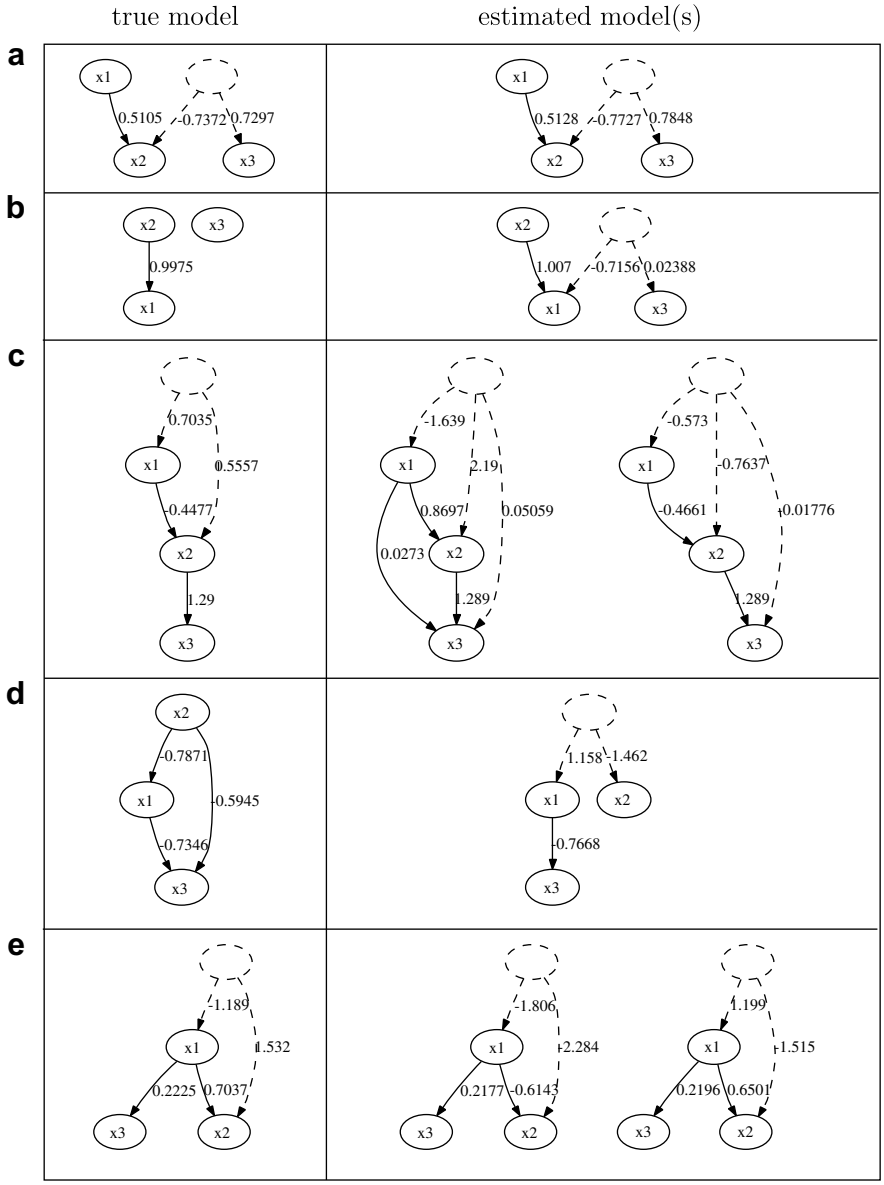

Fig. 6. Left column: original generating model. Right column: estimated set of models. See main text for details.

models was correct. However, note that we have nevertheless been able to determine the direction of causal influence: we know that x_1 affects x_2, and that x_2 affects x_3, and the strength of the latter effect is approximately 1.3.

In (d), we see a clear failure of the method. The estimation procedure has mistakenly selected a model with a hidden variable although there was none in the generating model. Furthermore, although the causal influence of x_1 on x_3 is correctly identified, the estimated model predicts that x_2 is causally unrelated to the other variables. The mistake is most likely due to the fact that the original model is close to being unfaithful. The indirect effect of x_2 on x_3 through x_1 is $(-0.7871) * (-0.7346) \approx 0.58$, which almost precisely cancels out the direct effect of x_2 on x_3, leaving the total effect very close to zero. For any reasonable sample sizes, this may confuse the method.

Finally, in (e) our method has again produced two observationally equivalent candidate models. Again, the rightmost is causally nearly identical to the true model, as the sign of the confounder is arbitrary (as discussed for (c) above). Since the two alternative models give very different predictions for the causal effect of x_1 on x_2 this effect cannot be determined from the data. On the contrary, there is unambiguous evidence for a positive effect of x_1 on x_3.

In total, we ran the model estimation algorithm a total of 128 times on datasets produced by models with two or three observed variables and zero or one hidden variables. Out of these, the model estimation succeeded (in the sense that all but one succeeded in Fig. 6) in a total of 98 cases, whereas 30 failed. When the estimation failed, it was practically always due to errors in the estimate of the ICA mixing matrix. Since overcomplete basis ICA algorithms are still an active research topic, it is likely that much better results can be obtained when more accurate methods are developed.

Alltogether, these simulations indicate that the basic method has potential, and that the learning of (small) latent variable LiNGAM models from data is feasible, even when the distributions involved are not known, nor is even the presence or absence of hidden variables.

6. Discussion

In this paper, we have shown that *in principle* it is possible to estimate, up to a finite set of observationally equivalent models, causal models involving linear relationships between non-Gaussian variables, some of which may be hidden. Additionally we have shown that this is possible even in practice, at least for very small networks. Nevertheless, there is still some way to go before we could seriously recommend the method as a practical data-analysis tool for real-world problems.

One important question is the possible presence of feedback. For example, in the motivating examples given in Section 2, it could well be argued that both x_1 causes x_2 *and* x_2 causes x_1. That is, we could have a feedback system. However, it could be argued that, if the feedback effects were not too fast, one could model the temporal dynamics explicitly, in which case directed acyclic models such as ours would be adequate. For example, for the poverty–illiteracy problem, one would separately represent GDP at time t and illiteracy at a sufficiently later time t', or vice versa, in which case feedback would not be an issue. For such data, one can obviously additionally rule out the models in which effects precede their causes (models III and VI, assuming that x_1 precedes x_2 in time), shortening the list of possible models but leaving many possibilities still. Thus, the framework presented in this paper can be applicable to feedback problems as long as one explicitly models the dynamical process. Specifically note that the possibility to detect (and take into account) unobserved common causes is something not available in standard time-series analysis.

Even when applying the method to problems which are guaranteed to be acyclical, there are a number of important issues to be solved with the present framework. Perhaps the most critical ones are the absence of measures of (a) how statistically reliable the result is, and (b) even if the result is reliable, how well the model actually fits the data. For instance, if the data is close to Gaussian, no algorithm will be able to reliably estimate an ICA basis from a reasonable number of datapoints, and the end result will necessarily be statistically unreliable. On the other hand, if the true model is non-linear, for example, it is quite possible that the result is consistent but nevertheless not useful at all. The question of statistical reliability can of course be solved using further resampling techniques. We employed bootstrapping only for the purposes of estimating a single model; one could equally well use the method to derive a number of models from which statistical reliability could be

estimated. But statistical reliability is only necessary, not sufficient, for the model to be a good description of the data. Thus, we need to evaluate the degree to which the model density fits the data, and be able to compare the present model with other explanations of the data. A Bayesian treatment of the model might be useful for this purpose. Another useful development would be to combine the present non-Gaussian framework with the standard Gaussian framework based on conditional independencies, so as to be able to combine the strengths of both methods. We are currently working in this direction.

We must also point out that the extensive use of resampling techniques comes at a significant computational cost. As the current algorithm for overcomplete ICA estimation is quite computationally expensive, the combination of these two family of methods has made it impossible to estimate anything but the smallest of networks. There is thus a strong need for either (a) faster but still accurate ICA estimation algorithms which can adapt to arbitrary distributions and overcomplete bases, and/or (b) analytical results on the degree of uncertainty in the ICA basis estimates. For the overcomplete ICA case, solutions to both of these problems are still very much under development.

One interesting possibility for learning larger models within a reasonable amount of time derives from the fact that larger networks may often have a modular structure which makes it possible to break them down into smaller ones. The key is that hidden variables only cause what might be called *local overcompleteness*. In a large lvLiNGAM model where each latent variable directly influences only a small number of observed variables, the data follows an independent subspace (ISA) model [9]. In such a model, the data distribution can be represented as a combination of low-dimensional independent subspaces. When the data follows the ISA model, the individual subspaces might still be found using efficient ICA algorithms [10], and algorithms for overcomplete ICA would only need to be run *inside each subspace*.

Finally, we want to acknowledge what is perhaps the weakest point of the framework developed in this paper. The method relies strongly on the assumption of linearity of the relationships between the measured variables. Linear models are, of course, widespread and very useful in most branches of science. Nevertheless, it is probably safe to say that the majority of interesting systems are to some extent non-linear, and it seems quite likely that the linear, non-Gaussian method developed here is particularly sensitive to non-linearities in the data. An empirical investigation of the sensitivity of the method to violations of linearity is an important future step.

7. Conclusions

We have recently shown how to estimate linear causal models when the distributions involved are non-Gaussian and no confounding hidden variables exist [14]. In this contribution, we have (a) given an introduction to this previous work, and (b) extended the method to take into account the effect of confounding latent variables. We have shown how to estimate the set of models consistent with the observed data; in some cases, this set consists of a single model, in which case all causal effects are identified. In all other cases, there is a *finite* set of observationally equivalent models of which only one is causally equivalent to the generating model. That is, we obtain a finite number of possible causal effects between any pair of observed variables. Simulations, for which full Matlab code is available for download, confirm the theoretical developments and demonstrate that small networks can indeed be estimated from data. Further work is still needed to develop the software into a practical, easy-to-use data-analysis tool.

Acknowledgements

We thank Aapo Hyvärinen for comments on the manuscript. P.O.H. was supported by the Academy of Finland Project #204826. S.S. would like to thank Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science and Transdisciplinary Research Integration Center, Research Organization of Information and Systems.

References

[1] C.A. Anderson, Violent video games: myths, facts, and unanswered questions, Psychological Science Agenda: Science Briefs 16 (5) (2003) 1–3.

[2] H. Attias, Independent factor analysis, Neural Computation 11 (1999) 803–851.
[3] K.A. Bollen, Structural Equations with Latent Variables, John Wiley & Sons, 1989.
[4] P. Comon, Independent component analysis – a new concept? Signal Processing 36 (1994) 287–314.
[5] G. Cooper, A Bayesian method for causal modeling and discovery under selection, in: Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, 2000, pp. 98–106.
[6] B. Efron, R. Tibshirani, Introduction to the Boostrap, Chapman and Hall, 1993.
[7] J. Eriksson, V. Koivunen, Identifiability, separability and uniqueness of linear ICA models, IEEE Signal Processing Letters 11 (7) (2004) 601–604.
[8] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (1969) 424–438.
[9] A. Hyvärinen, P.O. Hoyer, Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces, Neural Computation 12 (7) (2000) 1705–1720.
[10] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley Interscience, 2001.
[11] M.S. Lewicki, T.J. Sejnowski, Learning overcomplete representations, Neural Computation 12 (2) (2000) 337–365.
[12] E. Moulines, J.-F. Cardoso, E. Gassiat, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97), Munich, Germany, 1997, pp. 3617–3620.
[13] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000.
[14] S. Shimizu, P.O. Hoyer, A. Hyvärinen, A.J. Kerminen, A linear non-gaussian acyclic model for causal discovery, Journal of Machine Learning Research 7 (2006) 2003–2030.
[15] R. Silva, R. Scheines, C. Glymour, P. Spirtes, Learning the structure of linear latent variable models, Journal of Machine Learning Research 7 (2006) 191–246.
[16] P. Spirtes, Limits on causal inference from statistical data, in: American Economics Association Meeting, 1997.
[17] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, MIT Press, 2000.
[18] UNDP, Human Development Report, 2006. <http://hdr.undp.org/hdr2006/report.cfm>.