Finding a causal ordering via independent component analysis

Shohei Shimizu [a,b,1,2] Aapo Hyvärinen [a] Patrik O. Hoyer [a]
Yutaka Kano [b]

[a] *Helsinki Institute for Information Technology, Basic Research Unit, Department of Computer Science, University of Helsinki, Finland*

[b] *Division of Mathematical Science, Osaka University, Japan*

Abstract

The application of independent component analysis to discovery of a causal ordering between observed variables is studied. Path analysis is a widely-used method for causal analysis. It is of confirmatory nature and can provide statistical tests for assumed causal relations based on comparison of the implied covariance matrix with a sample covariance. However, it is based on the assumption of normality and only uses the covariance structure, which is why it has several problems, for example, one cannot find the causal direction between two variables if only those two variables are observed because the two models to be compared are equivalent to each other. A new statistical method for discovery of a causal ordering using non-normality of observed variables is developed to provide a partial solution to the problem.

Key words: Independent component analysis, non-normality, independence, causal inference, non-experimental data

[1] Corresponding author. Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan. Email: shimizu@sigmath.es.osaka-u.ac.jp.

[2] We thank Jarmo Hurri and Mika Inki for comments and interesting discussions. S.S. was supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture and Sports, Japan. A.H. was supported by the Academy of Finland through an Academy Research Fellow Position and project #203344. P.O.H. was supported by the Academy of Finland project #204826.

1 Introduction

An effective way to examine causality is to conduct an experiment with random assignment (Holland, 1986; Rubin, 1974). However, there are many situations that pose some difficulties to conduct experiments: One of the difficulties is that the direction of causality is often unknown. It is necessary to develop useful methods for finding a good initial model of causal orders between observed variables from non-experimental data.

Path analysis was originated by the biologist S. Wright in the 1920's and has been often applied to analyze causal relations of non-experimental data in an empirical way. The path analysis is an extension of regression analysis where many endogenous and exogenous variables can be analyzed simultaneously. In the 1970's, the path analysis was incorporated with factor analysis and latent variables were allowed in the model. The new framework is now called structural equation modeling (e.g., Bollen, 1989) and is a powerful tool of causal analysis.

However, the structural equation modeling (SEM) is of confirmatory nature and researchers have to model the true causal relationships based on background knowledge *before* collecting or analyzing data (e.g., Goldberger, 1972). It is difficult to model true causal relations in many cases, especially at the beginning of research. Lack of background knowledge often has the consequence that the causal direction is unknown.

Furthermore, SEM has some problems due to its restriction to normal distribution, for example: One cannot find the possible causal direction between two variables if only those two variables are observed because the two models with different direction are equivalent to each other.

A very simple illustration of the problem of finding the direction of causality is given by two regression models, called Model 1 and Model 2 here:

$$\text{Model 1:} \quad x_1 = b_{12}x_2 + \xi_1 \tag{1}$$
$$\text{Model 2:} \quad x_2 = b_{21}x_1 + \xi_2, \tag{2}$$

where the explanatory variable is assumed to be uncorrelated with the disturbance ξ_1 or ξ_2. We cannot say anything about which model is better from the two conventional regression analyses based on the two models above in the framework of SEM. Using the SEM terminology, the both models are saturated on the covariance matrix of $[x_1, x_2]$.

Kano and Shimizu (2003) and Shimizu and Kano (2003b) showed that use of non-normality of observed variables makes it possible to distinguish between

Model 1 and Model 2. In this paper, we shall extend their method to more than two variables and propose an algorithm to explore a causal ordering between observed variables from non-experimental data.

2 Brief review of independent component analysis

Independent component analysis (ICA) is a multivariate analysis technique which aims at separating or recovering linearly-mixed unobserved multidimensional independent signals from the mixed observable variables (Hyvärinen, Karhunen and Oja, 2001).

Let $\boldsymbol{x}$ be an observed m-vector. The ICA model for $\boldsymbol{x}$ is written as

$$\boldsymbol{x} = A\boldsymbol{s}, \tag{3}$$

where A is called a mixing matrix and $\boldsymbol{s}$ is an n-vector of unobserved variables or independent components with zero mean and unit variance. Typically, the number of observed variables m is assumed to equal that of latent variables n. The main goal of ICA is to estimate the mixing matrix.

Comon (1994) provided conditions for the model to be estimable for the typical case where $m \leq n$. The conditions include that the components of $\boldsymbol{s}$ are mutually independent and contain at most one normal component. ICA solves the estimation problem by maximizing independence among the components of $\boldsymbol{s}$. The independence is very often measured by non-normality (see e.g., Hyvärinen and Kano, 2003). That is, the estimation is implemented by finding the demixing matrix W such that the components of $\hat{\boldsymbol{s}} = W\boldsymbol{x}$ have maximal non-normality. A classical measure of non-normality is kurtosis, defined as

$$\mathrm{kurt}(u) = E(u^4) - 3\{E(u^2)\}^2. \tag{4}$$

The kurtosis is zero for a normal variable and non-zero for most non-normal variables. Comon (1994) proposed an estimation algorithm to maximize the sum of squared kurtosis of $\hat{\boldsymbol{s}}$, that is,

$$W = \arg\max_W \sum_{i=1}^{n} \mathrm{kurt}(\hat{s}_i)^2 = \arg\max_W \sum_{i=1}^{n} \mathrm{kurt}(\boldsymbol{w}_i^T \boldsymbol{x})^2, \tag{5}$$

where $\boldsymbol{w}_i^T$ denotes the i-th row of W. Here, the data is assumed to be sphered (whitened) (e.g., Hyvärinen, Karhunen and Oja, 2001) and W is constrained to be orthogonal.

Although the idea of ICA using kurtosis is simple, it can be very sensitive to outliers. Hyvärinen (1999) suggested a class of non-normality measures

$$J(u) \propto [E(G(u)) - E(G(\nu))]^2, \tag{6}$$

where $G(\cdot)$ is a nonlinear and nonquadratic function and ν follows the normal distribution with zero mean and unit variance. More robust estimators are obtained if G does not grow too fast. For example, one can take $G(u) = \log\cosh(u)$. He further proposed a very efficient algorithm to estimate W maximizing (6), called FastICA (Hyvärinen, 1999; Hyvärinen and Oja, 1997).

In ICA as well as the traditional multivariate methods including factor analysis, the following ambiguities hold: i) one cannot determine the sign of s_i. one can multiply the independent component by -1 without affecting the model in (3); ii) one cannot determine the order of the independent components. A permutation matrix P and its inverse can be substituted in the model to provide $\boldsymbol{x} = AP^{-1}P\boldsymbol{s}$. The element of $P\boldsymbol{s}$ are the original s_i, but in another order.

3 Finding a causal order between two variables

In this section, we shall explain how we can find a causal order between two variables using non-normality.

3.1 Definition of a causal order

What is causality? Many philosophers and statisticians have tried to answer the quite difficult question and proposed various frameworks to find causal relations for a long time (Bollen, 1989; Bullock, Harlow and Mulaik, 1994; Granger, 1969; Holland, 1986; Hume, 1740; Mill, 1843; Mulaik and James, 1995; Pearl, 2000; Rubin, 1974; Suppes, 1970).

In this article, we say that causality (a causal order) from a random variable x_1 to a random variable x_2, which we denote by $x_1 \to x_2$, is confirmed if an equation:

$$x_2 = f(x_1, \xi_2), \tag{7}$$

holds where ξ_2 is a disturbance variable which is distributed independently from the explanatory variable x_1.[3] The ξ_2 is a function of many variables $z_1, z_2, \cdots, z_q$ that have small and not very important influences on x_2 or that may not be noticed by researcher, as well as an error variable e_2. That is, $\xi_2 = g(z_1, z_2, \cdots, z_q, e_2)$ (e.g., Bollen, 1989).

For simplicity, let us assume that $f(x_1, \xi_2) = b_{21}x_1 + \xi_2$, is a simple linear function of x_1 and ξ_2. Then we obtain a simple regression analysis model:

$$x_2 = b_{21}x_1 + \xi_2, \tag{8}$$

where x_1 and ξ_2 are independent from each other. Now we can reformulate the causal order of x_1 to x_2: a nonzero constant b_{21} exists so that (8) holds. Note that independence between an explanatory variable x_1 and a disturbance variable ξ_2, not only their uncorrelatedness, is assumed here.[4]

The two concepts, independence and uncorrelatedness are very different. The independence between s_1 and s_2 is equivalent to

$$E[h_1(s_1)h_2(s_2)] - E[h_1(s_1)]E[h_2(s_2)] = 0. \tag{9}$$

for any two functions h_1 and h_2. Uncorrelatedness is a much weaker condition than independence. Two random variables s_1 and s_2 are said to be uncorrelated if their covariance is zero,

$$E(s_1 s_2) - E(s_1)E(s_2) = 0. \tag{10}$$

If those two variables are independent, they are uncorrelated, which follows directly from (9) taking $h_1(s_1) = s_1$ and $h_2(s_2) = s_2$. However, uncorrelatedness does not imply independence (see, e.g., Hyvärinen and Oja, 2000).

A dependence between x_1 and ξ_2 would imply the existence of one (or more) unobserved confounding variables between x_1 and x_2 (Bollen, 1989; Kano and Shimizu, 2003). It is known that regression-based causal analysis may be completely distorted if there are unobserved confounding variables. If x_1 and ξ_2

[3] Rigorously speaking, we need to examine if equation (7) holds for each unit in a population U to confirm causation from x_1 to x_2 in U because we have to distinguish between interpersonal change (causation) and individual difference (association) (see, e.g., Holland, 1986, for causation and association). However, it is rarely possible to examine it from non-experimental data since the data is usually one-time-point data. In this article, we assume that interpersonal change can be approximated by individual difference in our data sets, which is usually assumed in causal analysis based on non-experimental data.

[4] The condition is related to pseudo-isolation in Bollen (1989). However, he required only uncorrelatedness, not independence.

are independent, it implies that no unobserved confounding variable exists (Kano and Shimizu, 2003). However, if they are merely uncorrelated, it does not ensure anything about the existence of confounding variables. Let z be an unobserved confounding variable, and let us assume that

$$x_2 = b_{21}x_1 + \gamma_{23}z + \xi_2 \tag{11}$$

$$x_1 = \gamma_{13}z + \xi_1. \tag{12}$$

We then have

$$\mathrm{Cov}(x_1, x_2) = b_{21}\mathrm{Var}(x_1) + \gamma_{23}\gamma_{13}\mathrm{Var}(z). \tag{13}$$

Depending on the particular values of γ_{23} and γ_{13}, there could be nonzero covariance between x_1 and x_2 even if b_{21}=0. Using correlations alone, one could make an interpretation that a causal order from x_2 to x_1 or its opposite exists; on the other hand, there could be zero covariance between x_1 and x_2 even if b_{21} is large enough. Thus, independence and non-normality are key assumptions in our settings.

3.2 Finding a causal order between two variables

Let x_{1j} and x_{2j} $(j = 1, \ldots, N)$ be observations on random variables x_1 and x_2 with zero mean. Denote $\overline{x_i^2} = \frac{1}{N}\sum_{j=1}^{N} x_{ij}^2$ $(i = 1, 2)$ and $\overline{x_1x_2} = \frac{1}{N}\sum_{j=1}^{N} x_{1j}x_{2j}$. We shall use similar notation in subsequent derivations without explicit definitions.

The second-order moment structure of Model 1 is obviously given as

$$E\begin{bmatrix} \overline{x_1^2} \\ \overline{x_1x_2} \\ \overline{x_2^2} \end{bmatrix} = \begin{bmatrix} b_{12}^2E(x_2^2) + E(\xi_1^2) \\ b_{12}E(x_2^2) \\ E(x_2^2) \end{bmatrix} \quad \text{which we denote by} \quad E[\boldsymbol{m}_2] = \boldsymbol{\sigma}_2(\boldsymbol{\tau}_2),$$

where $\boldsymbol{\tau}_2 = [E(x_2^2),\ E(\xi_1^2),\ b_{12}]^T$. The number of sample moments to be used and the number of parameters $(E(x_2^2),\ E(\xi_1^2),\ b_{12})$ are both three and thus, the Models 1 and 2 are saturated and equivalent to each other as far as covariances alone are concerned. Both models receive a perfect fit to the sample covariance matrix.

Shimizu and Kano (2003b) assumed that $[x_1, x_2]$ is non-normally distributed and utilized higher-order moments to distinguish between Model 1 and Model

2. They further assumed that explanatory and disturbance variables, x_2 and ξ_1, x_1 and ξ_2, are independently distributed.

Consider using fourth-order moments. The expectations of the fourth-order moments can be expressed in a similar manner as

$$E\begin{bmatrix} \overline{x_1^4} \\ \overline{x_1^3 x_2} \\ \overline{x_1^2 x_2^2} \\ \overline{x_1 x_2^3} \\ \overline{x_2^4} \end{bmatrix} = \begin{bmatrix} b_{12}^4 E(x_2^4) + 6b_{12}^2 E(x_2^2)E(\xi_1^2) + E(\xi_1^4) \\ b_{12}^3 E(x_2^4) + 3b_{12} E(x_2^2)E(\xi_1^2) \\ b_{12}^2 E(x_2^4) + E(x_2^2)E(\xi_1^2) \\ b_{12} E(x_2^4) \\ E(x_2^4) \end{bmatrix}$$

which we denote by $E[\boldsymbol{m}_4] = \boldsymbol{\sigma}_4(\boldsymbol{\tau}_4)$,

for Model 1, where $\boldsymbol{\tau}_4 = [\boldsymbol{\tau}_2^T,\ E(x_2^4),\ E(\xi_1^4)]^T$.

In Model 1, we have three second-order moments and five fourth-order moments, whereas there are five parameters. The number of parameters is smaller than the number of moments used. Thus, if we define a measure of model fit by a weighted distance between the observed moments and the moments implied by the model as

$$T = N\left(\begin{bmatrix} \boldsymbol{m}_2 \\ \boldsymbol{m}_4 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\sigma}_2(\widehat{\boldsymbol{\tau}}_2) \\ \boldsymbol{\sigma}_4(\widehat{\boldsymbol{\tau}}_4) \end{bmatrix}\right)^T \hat{M} \left(\begin{bmatrix} \boldsymbol{m}_2 \\ \boldsymbol{m}_4 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\sigma}_2(\widehat{\boldsymbol{\tau}}_2) \\ \boldsymbol{\sigma}_4(\widehat{\boldsymbol{\tau}}_4) \end{bmatrix}\right), \tag{14}$$

with appropriate estimators $\widehat{\boldsymbol{\tau}}_i$ and a correctly chosen weight matrix $\hat{M}$, then T represents distance between data and the model employed and will be asymptotically distributed according to the chi-square distribution with df= 3 degrees of freedom. See Section 5 for some details. We can thus evaluate a fit of Model 1 using the statistic T. The same argument holds for Model 2, and we can confirm that Models 1 and 2 are not equivalent to each other in general, that is, the independence assumption between explanatory and disturbance variables is better fitted to one model than the other.

4 Finding a causal ordering between more than two variables based on ICA

In this section, we propose a new method of finding causal orders that generalizes our previous work, reviewed in the preceding section, to more than two variables.

4.1 Definition of a causal ordering

We say that observed variables x_i have a causal ordering if they can be ordered so that each variable is a function of the preceding variables plus an independent disturbance variable ξ_i. Let us denote this ordering by $i(1), \ldots, i(n)$.

In other words, we say that random variables, $x_1, x_2, \cdots, x_n$, have a causal ordering, $x_{i(1)} \to x_{i(2)} \to \cdots \to x_{i(n)}$, if nonzero coefficients $\beta_{i(j),i(k)}$ $(j = 1, 2, \cdots, n, k < j)$ exist so that the equations:

$$x_{i(j)} = \sum_{k=1}^{j-1} \beta_{i(j),i(k)} x_{i(k)} + \xi_{i(j)} \text{ for all } j = 1, \cdots, n, \tag{15}$$

hold where $\xi_{i(j)}$ is a disturbance variable and is independently distributed from $x_{i(k)}$ and from $\xi_{i(k)}$ for all $k < j$.

4.2 Definition of data model

Our definition of causality in (15) can also be interpreted as a data model. In the following, we actually assume that the data follows such a model so that the causal ordering is possible to find. Thus, we assume the following data model:

$$x_{i(j)} = \sum_{k=1}^{j-1} b_{i(j),i(k)} x_{i(k)} + \xi_{i(j)} \text{ for all } j = 1, \cdots, n. \tag{16}$$

We also assume that the disturbance variables $\xi_{i(j)}$ are non-normal, and mutually independent. This implies that $\xi_{i(j)}$ is independent from $x_{i(k)}$ for all $k < j$.

To investigate the causal structure of the x_i, we would like to find the correct ordering $i(j)$. Thus, the problem is finding the permutation of the observed variables that reflects the causal structure of the data. In what follows, we will show how such an ordering can be identified.

Let us normalize the equation (16) so that the disturbance variables ξ_i have unit variance. Denoting

$$w_{i(j),i(j)} = 1/\sqrt{\mathrm{var}(\xi_{i(j)})} \tag{17}$$

$$w_{i(j),i(k)} = -b_{i(j),i(k)}/\sqrt{\mathrm{var}(\xi_{i(j)})} \quad \text{for } k \neq j, \tag{18}$$

the equation (16) can be expressed as:

$$w_{i(j),i(j)} x_{i(j)} = \sum_{k<j} -w_{i(j),i(k)} x_{i(k)} + \xi^*_{i(j)}, \tag{19}$$

where $\xi^*_{i(j)}$ are the disturbance variables standardized to have unit variance.

Let us denote by $\tilde{\boldsymbol{x}}$ the vector where the observed variables are ordered according to $i(j)$. In matrix form, equation (16) can be expressed as

$$\tilde{\boldsymbol{x}} = B\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\xi}}, \tag{20}$$

where the matrix B is *lower triangular*. Using W, this becomes

$$\mathrm{diag}(W)\tilde{\boldsymbol{x}} = -\mathrm{offdiag}(W)\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\xi}}^* \quad \text{or equivalently} \quad W\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{\xi}}^*, \tag{21}$$

where W is still lower triangular, for the correct permutation of the observed variables. This corresponds to the correct permutation of the columns of W. From the theory of ICA, we know that this W can be estimated up to a permutation of its rows, using standard ICA methods.

Now we can use the following theorem:

Theorem 1 *If W is lower triangular and all the elements w_{ij} are nonzero for $i \geq j$, no other permutation of rows and columns is lower triangular*

Proof First, note that any joint permutation of rows and columns can be performed by first permuting the rows and then the columns. This is because the permutations of rows or columns can be expressed by left and right multiplication by permutation matrices, respectively, and any product of multiple permutations therefore reduces to a multiplication by two permutation matrices, one from the right and one from the left, and either of the multiplications can be done first. Assume a permutation of rows has been done, and denote this new matrix by $W^\dagger$. Assume that the first row in $W^\dagger$ is not the same as the first row in W. Then, at least two elements on the first row of $W^\dagger$ are nonzero. Now, any permutation of columns cannot change the number of nonzero elements on the first row. Thus, a combination of row and column permutation that is lower-triangular must be such that the first row of the row-permuted matrix $W^\dagger$ is equal to the first row of W. Also, the column-permutation cannot move the first column in order to preserve lower-triangularity. Thus, we have proven that the first row must remain the first row, and the first column must remain the first column. The same proof can be applied on every row and column in succession, which proves the theorem.

Therefore, if the $b_{i(j),i(k)}$ are not zeros, the permutation to make

$$W = \mathrm{Var}(\boldsymbol{\xi})^{-1/2}(I_n - B), \tag{22}$$

lower triangular is unique. The I_n denotes an n-dimensional identity matrix. Then the causal ordering between x_i is uniquely determined taking the $b_{i(j),i(k)}$ as the $\beta_{i(j),i(k)}$ in (15).

In practice, because of finite sample effects, all elements of the W given by ICA are non-zero even if the model holds. That is, those entries which should be exactly zero are only approximately zero. This complicates the search for the correct row and column permutations, requiring us to search for permutations which yield *approximate* lower-triangularity.

Here, we suggest to optimize lower triangularity using a cost function which sums the squares of all entries above the main diagonal, i.e.

$$C = \sum_{j<k} w_{jk}^2. \tag{23}$$

Finding the globally optimal row and column permutations for this problem is hard. Our approach is to employ a *greedy* column- and row-swapping algorithm: At each step, calculate the change in C resulting from every possible pairwise swap of columns or swap of rows, and perform the swap which decreases C the most. This step is iterated until no pairwise swap can decrease the cost any further.

Although the above-described greedy algorithm often gets stuck in suboptimal local minima, it is very fast to run. This implies that one can run it from a large number of different random initial permutations in a reasonable time, leading to a very high probability of finding the global optimum. A Matlab implementation of this permutation algorithm is available at `http://www.cs.helsinki.fi/u/phoyer/code/csdapack.tar`.

If we have correctly permuted W, the disturbance standard deviation $\sqrt{\mathrm{var}(\xi_{i(j)})}$ can be estimated by $1/w_{i(j),i(j)}$ from (17). [5] Then we obtain the estimate of B by

$$\hat{B} = I_n - \mathrm{diag}(\hat{W})^{-1}\hat{W}. \tag{24}$$

Thus, the model (20) can be estimated by

(1) estimating an initial W by ICA,
(2) finding a combination of permutations of the rows and the columns of $\tilde{W}$ so that $\tilde{W}$ becomes as close to lower triangular as possible, using the algorithm above,
(3) estimating B by $I_n - \mathrm{diag}(\tilde{W})^{-1}\tilde{W}$.

The $\tilde{W}$ denotes a correctly permuted version of W. It should be noted that the correct causal ordering is given by the permutation of the columns found by our method. The correct permutation of rows and the value of B are additional information that are not always necessary.

4.4 Example

Now we shall show the models 1 and 2 can be expressed in this framework. In Model 1, the causal order of observed variables is $(i(1), i(2)) = (2, 1)$. Model 1 can be rewritten as:

$$\text{Model 1:} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & b_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \tag{25}$$

$$\Leftrightarrow \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b_{12} & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} \xi_2 \\ \xi_1 \end{bmatrix}. \tag{26}$$

Here $\tilde{\boldsymbol{x}}$ and B are

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ b_{12} & 0 \end{bmatrix}. \tag{27}$$

[5] ICA has the sign ambiguity. The estimated w_{ii} could have a negative sign. Then we multiply the i-th row vector $\boldsymbol{w}_i^T$ by -1 so that w_{ii} has a positive sign.

One can see that the B is lower triangular when the observed variables are ordered according to $i(j)$. Also in Model 2, one can see the lower triangularity of B in the same manner.

4.5 Alternative approach

Above, we said that the causal ordering between observed variables is uniquely determined if all the $b_{i(j),i(k)}$ are not zeros, which is a sufficient but not a necessary condition. There is another possibility where the causal ordering is unique. Let A be the inverse of W. Note that A is also lower triangular. The a_{ij}/a_{jj} represents the total effect of x_j to x_i, whereas the b_{ij} the direct effect of x_j to x_i (see, Bollen, 1989, for total effect and direct effect). Then the model (20) can be rewritten as

$$\tilde{\boldsymbol{x}} = A\tilde{\boldsymbol{\xi}}^*, \tag{28}$$

which is the ICA model in (3) and the A is estimable up to a permutation of its columns, using standard ICA methods. We can find the optimal permutations in A in the same manner as finding those in W. Now, the causal ordering between x_i is uniquely determined if $a_{i(j),i(k)}$ are not zeros (Theorem 1).

The link between the lower triangularity of A and the causal ordering can be seen as follows. For the lower triangular mixing matrix, $x_{i(1)}$ is essentially equal to $\xi^*_{i(1)}$, up to a multiplicative constant, $a_{i(1),i(1)}$. On the other hand, $x_{i(2)}$ is a function of $\xi^*_{i(1)}$ and $\xi^*_{i(2)}$, $a_{i(2),i(1)}\xi^*_{i(1)} + a_{i(2),i(2)}\xi^*_{i(2)}$. Thus, $x_{i(2)}$ is a function of $x_{i(1)}$ and a new independent variable, $\xi^*_{i(2)}$, that is, $(a_{i(2),i(1)}/a_{i(1),i(1)})x_{i(1)} + a_{i(2),i(2)}\xi^*_{i(2)}$. This indicates that $x_{i(1)}$ may cause $x_{i(2)}$, but $x_{i(2)}$ cannot cause $x_{i(1)}$. Continuing the same logic, we see that $x_{i(1)}$ can cause $x_{i(3)}$ and $x_{i(2)}$ can cause $x_{i(3)}$, but $x_{i(3)}$ cannot cause either $x_{i(1)}$ or $x_{i(2)}$ because $x_{i(3)}$ is simply a function of $x_{i(1)}$ and $x_{i(2)}$. In general, $x_{i(j)}$ is a function of $x_{i(1)}, \cdots, x_{i(j-1)}$ and $\xi^*_{i(j)}$, which establishes the direction of possible causality.

The two methods: i) W-based method; ii) A-based method compensate each other. For example, let us assume that

$$\begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix} \begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} + \begin{bmatrix} \xi_{i(1)} \\ \xi_{i(2)} \\ \xi_{i(3)} \end{bmatrix}, \tag{29}$$

where b_{21} and b_{32} are not zeros. The b_{31} is zero and the causal ordering may not be unique if the W-based method is applied. However, let us rewrite (29) as

$$\begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathrm{Var}(\xi_{i(1)}) & 0 & 0 \\ 0 & \mathrm{Var}(\xi_{i(2)}) & 0 \\ 0 & 0 & \mathrm{Var}(\xi_{i(3)}) \end{bmatrix}^{1/2} \begin{bmatrix} \xi^*_{i(1)} \\ \xi^*_{i(2)} \\ \xi^*_{i(3)} \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}(\xi_{i(1)})^{1/2} & 0 & 0 \\ b_{21}\mathrm{Var}(\xi_{i(1)})^{1/2} & \mathrm{Var}(\xi_{i(2)})^{1/2} & 0 \\ b_{32}b_{21}\mathrm{Var}(\xi_{i(1)})^{1/2} & b_{32}\mathrm{Var}(\xi_{i(2)})^{1/2} & \mathrm{Var}(\xi_{i(3)})^{1/2} \end{bmatrix} \begin{bmatrix} \xi^*_{i(1)} \\ \xi^*_{i(2)} \\ \xi^*_{i(3)} \end{bmatrix}. \qquad (30)$$

Then the causal ordering can be recovered by the A-based method.

Another simple example is:

$$\begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ b_{21} & 0 & 0 \\ b_{31} & b_{32} & 0 \end{bmatrix} \begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} + \begin{bmatrix} \xi_{i(1)} \\ \xi_{i(2)} \\ \xi_{i(3)} \end{bmatrix}, \qquad (31)$$

where b_{21}, b_{31}, b_{32} are not zeros and $b_{31} + b_{32}b_{21}$ is zero, for example, $b_{21} = 0.3, b_{31} = 0.6, b_{32} = -0.2$. Then all the direct effect of $x_{i(k)}$ to $x_{i(j)}$, $b_{i(j),i(k)}$ $(k < j)$, are not zeros and the causal ordering can be recovered by the W-based method. However, the A-based method fails to recover the causal ordering because the total effect of $x_{i(1)}$ to $x_{i(3)}$, $b_{31} + b_{32}b_{21}$, is zero:

$$\begin{bmatrix} x_{i(1)} \\ x_{i(2)} \\ x_{i(3)} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(\xi_{i(1)})^{1/2} & 0 & 0 \\ b_{21}\mathrm{Var}(\xi_{i(1)})^{1/2} & \mathrm{Var}(\xi_{i(2)})^{1/2} & 0 \\ (b_{31} + b_{32}b_{21})\mathrm{Var}(\xi_{i(1)})^{1/2} \; (= 0) & b_{32}\mathrm{Var}(\xi_{i(2)})^{1/2} & \mathrm{Var}(\xi_{i(3)})^{1/2} \end{bmatrix} \begin{bmatrix} \xi^*_{i(1)} \\ \xi^*_{i(2)} \\ \xi^*_{i(3)} \end{bmatrix}.$$

Thus both W-based and A-based methods are useful for finding a causal ordering between observed variables. In the latter part of this article, we report the simulation experiment and real example using only the W-based method to save space.

5 Examination of independence

In our setting, the independence assumption between explanatory and disturbance variables is crucial. We propose a test statistic to examine the independence assumption statistically.

Let N be a sample size and define V as

$$V = \lim_{N\to\infty} N \times \mathrm{Var}[\boldsymbol{m}_2^T,\ \boldsymbol{m}_4^T]^T. \tag{32}$$

Letting $\boldsymbol{\tau}$ be a vector that contains the model parameters and $\boldsymbol{m}_2$ and $\boldsymbol{m}_4$ be the vectorized second- and fourth-order moments after removing the redundant elements and $\boldsymbol{\sigma}_2(\boldsymbol{\tau}) = E(\boldsymbol{m}_2)$, $\boldsymbol{\sigma}_4(\boldsymbol{\tau}) = E(\boldsymbol{m}_4)$, the test statistic T to examine the model assumption is defined as

$$T = N\left(\begin{bmatrix}\boldsymbol{m}_2\\ \boldsymbol{m}_4\end{bmatrix} - \begin{bmatrix}\boldsymbol{\sigma}_2(\hat{\boldsymbol{\tau}})\\ \boldsymbol{\sigma}_4(\hat{\boldsymbol{\tau}})\end{bmatrix}\right)^T \hat{M}\left(\begin{bmatrix}\boldsymbol{m}_2\\ \boldsymbol{m}_4\end{bmatrix} - \begin{bmatrix}\boldsymbol{\sigma}_2(\hat{\boldsymbol{\tau}})\\ \boldsymbol{\sigma}_4(\hat{\boldsymbol{\tau}})\end{bmatrix}\right), \tag{33}$$

with

$$\hat{M} = \hat{V}^{-1} - \hat{V}^{-1}\hat{J}(\hat{J}^T\hat{V}^{-1}\hat{J})^{-1}\hat{J}^T\hat{V}^{-1}, \tag{34}$$

where

$$\hat{J} = \left.\frac{\partial[\boldsymbol{\sigma}_2(\boldsymbol{\tau})^T, \boldsymbol{\sigma}_4(\boldsymbol{\tau})^T]^T}{\partial\boldsymbol{\tau}^T}\right|_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}}}. \tag{35}$$

The statistic T approximates to a chi-square variate with degrees $\mathrm{tr}[VM]$ of freedom where N is large enough (e.g., Shimizu and Kano, 2003a). The required assumption for this is that $\hat{\boldsymbol{\tau}}$ is a $\sqrt{N}$-consistent estimator. No asymptotic normality is needed. See Browne (1984) for details.

6 Simulation experiments

We conducted simulation experiments to study the performance of the method developed above.

The simulation consisted of 1000 causal ordering recovery trials for data of two different dimensions: 3 and 15 variables x_i. In each trial, we generated three- or fifteen-dimensional random vector $\tilde{\boldsymbol{\xi}}^*$ of different sample sizes (see Table 1 and 2 below) as standardized disturbance variables where their components are independently distributed according to the t distribution with parameters yielding kurtoses from 2 to 6. The variables were standardized to have zero mean and unit variance.

A random lower triangular matrix B where the element b_{ij} $(i > j)$ was distributed according to the uniform distribution $U(0.2, 1)$ and multiplied by -1 with probability 50% was created. A random diagonal matrix D whose diagonal elements are independently distributed according to the uniform distribution $U(0.2, 1)$ was created. Then a random mixing matrix $A = (I-B)^{-1}D$ was computed. The standardized disturbance variables were linearly mixed by A after its rows were permuted randomly.

We employed FastICA [6] as an ICA method and took $\log\cosh(u)$ as $G(u)$ in (6), where the symmetric orthogonalization was applied (Hyvärinen, 1999; Hyvärinen and Oja, 1997).

The W-based method developed above was then applied on the data. We computed how many trials recovered the correct permutation of observed variables. We also compared the performance of our permutation algorithm to that of a bruteforce permutation algorithm that tried all the combinations of rows and columns of estimated W for three variables. (The bruteforce permutation algorithm is computationally too hard to apply to the fifteen variables cases.)

The results are shown in Table 1 and 2. The numbers of trials where FastICA did not converge were given in parentheses (if it occurred), and such trials were excluded when counting the numbers of successful recoveries. For three variables (Table. 1), more than 95% of causal orderings were recovered when $N = 500$. The table also implied that our permutation algorithm provided as good performance as the bruteforce permutation algorithm. For fifteen variables (Table. 2), more than 95% of causal orderings were recovered when $N = 2500$. Overall, we would say that our method successfully recovered the correct causal orderings for reasonable sample sizes.

Table 1
Numbers of recovered causal orderings for 3 variables out of 1000 trials. (The numbers of trials where FastICA did not converge are given in parentheses if it occurred.)

	N				
	100	250	500	750	1000
Our permutation alg.	509 (179)	813 (51)	963 (4)	977 (1)	990
Bruteforce permutation alg.	509 (179)	813 (51)	963 (4)	977 (1)	990

[6] The MATLAB package is available at http://www.cis.hut.fi/projects/ica/fastica/.

Table 2
Numbers of recovered causal orderings for 15 variables out of 1000 trials. (The numbers of trials where FastICA did not converge are given in parentheses if it occurred.)

	N				
	1000	1500	2000	2500	3000
Our permutation alg.	666 (6)	843	928	965	975

7 Real data example

Questionnaire data about criminal psychology were analyzed as an example to illustrate the effectiveness of our method. The survey was conducted to students at Osaka University, Japan (Murakami, 2000). The sample size was 222. Observed variables were standardized to have zero means and unit variances.

We explored a possible causal ordering between observed variables x_1, x_2 and x_3 using the W-based method proposed in Section 4. The labels of the variables x_1, x_2 and x_3 are shown in Table 3. According to a criminal psychology theory (Gottfredson and Hirschi, 1990), the frequency of criminal opportunities (x_2) is a typical environmental cause of the frequency of criminal behaviors (x_1) (Murakami, 2000). Also, the x_1 and x_2 were preceding in time to x_3. Therefore, the possible causal ordering from the background knowledge was $x_2 \rightarrow x_1 \rightarrow x_3$. However, these were of course unknown to our method. The aim in this real example was to know if our method was really able to find the correct causal ordering without any background knowledge.

The kurtoses of the variables were 1.29, 8.71, 2.35 respectively and a Kolmogorov-Smirnov test showed that all variables could not be assumed to come from the normal distribution (significance level 1%). Thus, statistical methods based on the non-normal assumption including our method should be applicable on this kind of non-normal data.

Table 3
Variable labels

x_1: Sum of items that ask subjective evaluation on frequency of your criminal behavior when you went to high school
x_2: Sum of items that ask subjective evaluation on frequency of your criminal opportunities when you went to high school
x_3: Sum of items that ask subjective evaluation on frequency of your criminal behavior last one year

We employed FastICA, where $\log\cosh(u)$ is taken as $G(u)$ in (6) and the symmetric orthogonalization was applied.

The estimated W by FastICA was

$$\begin{bmatrix} -1.26 & 1.00 & -0.11 \\ 0.07 & 0.83 & 0.24 \\ 0.55 & 0.35 & -1.14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \xi_1^* \\ \xi_2^* \\ \xi_3^* \end{bmatrix}, \tag{36}$$

and the permuted W so that it becomes as lower triangular as possible was

$$\begin{bmatrix} 0.83 & 0.07 & 0.24 \\ 1.00 & -1.26 & -0.11 \\ 0.35 & 0.55 & -1.14 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} \xi_2^* \\ \xi_1^* \\ \xi_3^* \end{bmatrix}, \tag{37}$$

where the first and second rows and the first and second columns were permuted, respectively. The independence assumption was not rejected (T in (33) was 13.12 with p value of 0.16), which implied that no unobserved confounding variables existed. The result implied the causal ordering, $x_2 \rightarrow x_1 \rightarrow x_3$, that is, criminal opportunities at high schools $\rightarrow$ criminal behaviors at high schools $\rightarrow$ criminal behaviors last one year. The order $x_2 \rightarrow x_1$ was reasonable to the criminal psychology theory, and the order $x_1 \rightarrow x_3$ was reasonable to the time order. Therefore, the causal ordering founded by our method would be reasonable to the background knowledge.

Further, we conducted a simulation to study the stability of our method in this example. We generated 1000 data sets in the same manner as the simulations in the previous section other than two points: i) we created three standardized disturbance variables that had exactly the distributions of the standardized disturbance variables (independent components) found in the real example, which was possible simply by taking the independent components from the real data, and making them really independent reordering randomly each variable in the sample. (The kurtoses of the independent components found in the real data were 7.39, 2.16, 2.53.); ii) we computed *constant* matrices B and D:

$$B = \begin{bmatrix} 0 & 0 & 0 \\ 0.79 & 0 & 0 \\ 0.30 & 0.48 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1.21 & 0 & 0 \\ 0 & 0.79 & 0 \\ 0 & 0 & 0.87 \end{bmatrix}, \tag{38}$$

by setting the upper triangular elements of the $\tilde{W}$:

$$\tilde{W} = \begin{bmatrix} 0.83 & 0.07 & 0.24 \\ -1.00 & 1.26 & 0.11 \\ -0.35 & -0.55 & 1.14 \end{bmatrix}, \quad (39)$$

to zeroes and created a mixing matrix $A = (I - B)^{-1}D$.

Our method recovered 96.60 % of the causal orderings. The stability was quite good, and the causal ordering found was reasonable to the background knowledge. Therefore, we would conclude that our method successfully recovered the correct causal ordering in this real example.

8 Discussion

We developed a new statistical method for discovering a possible causal ordering using non-normality of observed variables. Whereas there are some approaches to causal analysis such as SEM, our approach based on ICA is totally different from them. SEM cannot find the direction of causality in many cases without much background knowledge because the normal assumption on SEM limits its applicability. We provided a partial solution to the problem utilizing non-normality of observed variables.

There are some drawbacks of our model. When the distribution is close to the normal distribution, our method is unstable. Linearity assumption is rather restrictive.

Researchers should and can make further confirmatory causal inferences including experimental and longitudinal studies based on the result of our exploratory causal inference method. The method developed here would be helpful to construct a good initial model.

References

Bollen, K. A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Bullock, H. E., Harlow, L. L. and Mulaik, S. A., 1994. Causal issues in structural equation modeling research. Structural Equation Modeling 1, 253-267.

Browne, M. W., 1984. Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology 9, 665-672.

Comon, P., 1994. Independent component analysis. A new concept?. Signal Processing 36, 62-83.
Goldberger, A. S., 1972. Structural equation models in the social sciences. Econometrica 40, 979-1001.
Gottfredson, M. R. and Hirschi, T., 1990. A general theory of crime. Stanford University Press, Stanford, CA.
Granger, C. W. J., 1969. Investigating causal relations by econometric models and cross-spectral method. Econometrica 37, 424-438.
Holland, P. W., 1986. Statistics and causal inference (with discussions). Journal of the American Statistical Association 81, 945-970.
Hume, D. A., 1740. Treatise on Human Nature. Dutton, New York.
Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transaction on Neural Networks 10, 626-634.
Hyvärinen, A. and Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483-1492.
Hyvärinen, A. and Oja, E., 2000. Independent component analysis: Algorithms and applications. Neural Networks 13, 411-430.
Hyvärinen, A., Karhunen, J. and Oja, E., 2001. Independent component analysis, Wiley, New York.
Hyvärinen, A. and Kano, Y., 2003. Independent component analysis for non-normal factor analysis. In: Yanai, H. et al. (Ed.), New Developments in Psychometrics, Springer Verlag, Tokyo, 649-656.
Kano, Y. and Shimizu, S., 2003. Causal inference using non-normality. In: Higuchi, T., Iba, Y. & Ishiguro, M., (Ed.), Proceedings of the International Symposium on Science of modeling -The 30th Anniversary of the Information Criterion (AIC)-, ISM Report on Research and Education, No.17, The Institute of Statistical Mathematics, Tokyo, 261-270.
Mill, J. S., 1843. A system of logic. Longmans, London.
Mulaik, S. A. and James, L. R., 1995. Objectivity and reasoning in science and structural equation modeling. In: Hoyle, H. (Ed.), Structural Equation Modeling, Sage Publications, CA, 118-137.
Murakami, N., 2000. Research on causes of criminal and deviant behavior. Bachelor thesis, School of Human Sciences, Osaka University. (In Japanese).
Pearl, J., 2000. Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge.
Rubin, D. B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688-701.
Shimizu, S. and Kano, Y., 2003a. Examination of independence in independent component analysis. In: Yanai, H. et al. (Ed.), New Developments in Psychometrics, Springer Verlag, Tokyo, 665-672.
Shimizu, S. and Kano, Y., 2003b. Non-normal structural equation modeling. (submitted).
Suppes, P. C., 1970. A Probabilistic Theory of Causality, North-Holland, Amsterdam.