Estimation of linear non-Gaussian acyclic models for latent factors

Shohei Shimizu[a] Patrik O. Hoyer[b] Aapo Hyvärinen[b,c]

[a] *The Institute of Scientific and Industrial Research, Osaka University*
Mihogaoka 8-1, Ibaraki, Osaka 567-0047, JAPAN

[b] *Dept. of Computer Science and Helsinki Institute for Information Technology,*
University of Helsinki, FIN-00014, Finland

[c] *Dept. of Department of Mathematics and Statistics, University of Helsinki,*
FIN-00014, Finland

Abstract

Many methods have been proposed for discovery of causal relations among observed variables. But one often wants to discover causal relations among latent factors rather than observed variables. Some methods have been proposed to estimate linear acyclic models for latent factors that are measured by observed variables. However, most of the methods use data covariance structure alone for model identification, and this leads to a number of indistinguishable models. In this paper, we show that a linear acyclic model for latent factors is identifiable when the data are non-Gaussian.

Key words: Independent component analysis, causal analysis, latent factors

1 Introduction

Many algorithms have been proposed for discovery of causal relations among observed variables [8,12]. But many empirical researchers are more interested in discovering causal relations among latent factors [11]. Some methods have been proposed for estimating linear structures of latent factors that are linearly measured by observed variables [11]. However, most of the methods employ covariance structure of data alone for model identification even when the data are non-Gaussian, and this leads to a number of indistinguishable models for latent factors. In this paper, we show that a linear acyclic model for latent factors is uniquely identified using non-Gaussian structures of data as well.

2 Background

In [10], a non-Gaussian variant of Bayesian networks was proposed. Assume that we observe data generated from a process with the following properties; (a) The generating process can be represented graphically by a DAG. (b) The relations between variables are linear. Without loss of generality, each observed variable x_i is assumed to have zero mean. Then we have

$$\boldsymbol{x} = \bar{\mathbf{B}}\boldsymbol{x} + \boldsymbol{e}, \tag{1}$$

where $\bar{\mathbf{B}}$ is a matrix that contains the connection strength from x_j to x_i denoted by $\bar{b}_{ij}$ and that could be permuted by simultaneous equal row and column permutations to strict lower triangularity due to the acyclicity assumption. (Strict lower triangularity is here defined as lower triangular with all zeros on the diagonal.); (c) The external influences e_i are all continuous random variables having *non-Gaussian* distributions with zero means and non-zero variances, and the e_i are independent of each other. A model with these three properties is called a *Linear, Non-Gaussian, Acyclic Model* (LiNGAM).

Now let us see how one can identify the parameters of the model. Solving Eq. (1) for $\boldsymbol{x}$ one obtains

$$\boldsymbol{x} = \bar{\mathbf{A}}\boldsymbol{e}, \tag{2}$$

where $\bar{\mathbf{A}} = (\mathbf{I} - \bar{\mathbf{B}})^{-1}$. Since the components of $\boldsymbol{e}$ are independent and non-Gaussian, Eq. (2) defines the standard linear independent component analysis (ICA) model [5], which is known to be identifiable [1].

While ICA is essentially able to estimate $\bar{\mathbf{A}}$ (and $\bar{\mathbf{W}} = \bar{\mathbf{A}}^{-1} = \mathbf{I} - \bar{\mathbf{B}}$), there are permutation and scaling indeterminacies. Therefore, ICA actually gives $\bar{\mathbf{W}}_{ICA}$=$\bar{\mathbf{P}}\bar{\mathbf{D}}\bar{\mathbf{W}}$, where $\bar{\mathbf{P}}$ is an *unknown* permutation matrix, and $\bar{\mathbf{D}}$ is an *unknown* diagonal scaling matrix. But in LiNGAM, one can find the correct permutation as described in [10]: the correct permutation is the only one which has no zeros in the diagonal. Further, one can find the correct scaling of the independent components. One only has to re-normalize the rows of $\bar{\mathbf{D}}\bar{\mathbf{W}}$ so that all the diagonal elements equal unity, which gives $\bar{\mathbf{W}}$. Then one can finally compute the connection strength matrix for observed variables $\bar{\mathbf{B}} = \mathbf{I} - \bar{\mathbf{W}}$.

3 A linear non-Gaussian acyclic model for latent factors

Let us consider the following linear model:

$$\begin{aligned} \boldsymbol{f} &= \mathbf{B}\boldsymbol{f} + \boldsymbol{d} \qquad (3)\\ \boldsymbol{x} &= \mathbf{G}\boldsymbol{f} + \boldsymbol{e}, \qquad (4) \end{aligned}$$

where $\boldsymbol{f}$, $\boldsymbol{d}$, $\boldsymbol{x}$ and $\boldsymbol{e}$ are random vectors that collect latent factors f_i, external influences d_i, observed variables x_i and errors e_i respectively, and $\mathbf{B}$ and $\mathbf{G}$ are matrices that collect connection strengths b_{ij} and factor loadings g_{ij}. We assume that the relations among latent factors f_i can be represented by a DAG, that is, $\mathbf{B}$ can be permuted to be strictly lower-triangular. The $\boldsymbol{e}$ and $\boldsymbol{f}$ are independent. The e_i and d_i are mutually independent, but f_i are allowed to be dependent due to the DAG structure. Without loss of generality, we assume that f_i are of zero mean and unit variance.

We now make some key assumptions of our approach; i) the Faithfulness[1] [12]; ii) each latent factor f_i has at least three *pure* measurement variables. Pure measurement variables are defined as observed variables that have a single latent factor parent [11]; iii) the external influences d_i are continuous random variables that follow non-Gaussian distributions whose means are zeros and moments exist. Note that the errors e_i can be either Gaussian or non-Gaussian. See Fig. 1 for a graphical example of our data generating models.

===== Insert Figure 1 =====

4 Model identification

First let us consider how to estimate the measurement model (4). In [11], the authors proposed a discovery algorithm called *BuildPureClusters* algorithm that returns the number of latent factors and which observed variable purely measures which latent factor in the model (4). Once the number of latent factors and their (at least 3) pure measurement variables are known, one can identify the factor loading matrix $\mathbf{G}$ and covariance matrices of factors and errors ($\text{cov}(\boldsymbol{f})$ and $\text{cov}(\boldsymbol{e})$) by factor analysis that constrains the pure measurement variables to be pure [9]. For example, in Fig. 1, the BuildPureClusters algorithm removes impure variables x_4 and x_8 and find three pure clusters i) x_1, x_2, x_3, ii) x_5, x_6, x_7, and iii) x_9, x_{10}, x_{11} that are measured by a single factor. Then we analyze all the observed variables using factor analysis with three factors constraining such coefficients in $\mathbf{G}$ to be zeros that are from each factor to observed variables the other factors purely measure. Note that variances of factors $\text{var}(f_i)$ are fixed to be unity and are not estimated, and covariances of errors $\text{cov}(e_i, e_j)$ $(i \neq j)$ are zeros due to the model assumption.

Next, we want to estimate the structural model (3). Using the relation $\boldsymbol{f} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{d}$ from (3), the reduced form of (3) and (4) is written as

[1] In the context, the faithfulness means that no combined effects of multiple pathways cancel out to be zeros and accidentally make correlations and partial correlations of variables equal to zeros [11].

$$\begin{aligned} \boldsymbol{x} &= \mathbf{G}(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{d}+\boldsymbol{e} \\ &= \mathbf{A}\boldsymbol{d}+\boldsymbol{e}, \end{aligned} \tag{5}$$

where $\mathbf{A} = \mathbf{G}(\mathbf{I}-\mathbf{B})^{-1}$. Since external influences d_i are non-Gaussian and independent, this is an ICA model with additive errors (Noisy ICA), which is identifiable up to an arbitrary permutation and scaling of columns of $\mathbf{A}$ [2,3].

Since the orders of columns of $\mathbf{G}$ would be different from $\mathbf{A}$ due to the permutation indeterminacy of factor analysis and noisy ICA, we obtain $\mathbf{W}$=$\mathbf{I}-\mathbf{B}$ with a random row permutation by $(\mathbf{G}^T\mathbf{A})^{-1}(\mathbf{G}^T\mathbf{G})$ using the relation $\mathbf{A} = \mathbf{G}(\mathbf{I}-\mathbf{B})^{-1}$. Further, there is the scaling indeterminacy of columns of $\mathbf{A}$ or rows of $\mathbf{W}$. That is, we obtain $\mathbf{W}_{noisyICA}$=$\mathbf{PDW}$, where $\mathbf{P}$ is an unknown permutation matrix, and $\mathbf{D}$ is an unknown diagonal scaling matrix. Fortunately, we can fix the permutation and scaling and obtain $\mathbf{W}$ in the same manner as the LiNGAM method described in Section 2. Thus, we can obtain the connection strength matrix for latent factors $\mathbf{B} = \mathbf{I}-\mathbf{W}$.

Existing estimation methods [6,7] for the noisy ICA model (5) with errors that follow arbitrary distributions are not computationally very feasible especially for large dimensions [5]. Therefore, we here use an approximate solution. We first perform principal component analysis (PCA) and subsequently apply ordinary noise-free ICA (FastICA [4] here) to the principal components. This is validated when the measurement errors are small enough *or* a large number of measurement variables for each factor are available [5].

5 Estimation procedure

We now propose an estimation procedure for the model in Section 3:

(1) Find the number of latent factors and which observed variable purely measures which latent factor by the BuildPureClusters algorithm.
(2) Estimate $\mathbf{G}$ by factor analysis constraining pure x_i found in Step (1) to be pure.
(3) Estimate $\mathbf{A}$ by noisy ICA. (Here we use PCA+FastICA.)
(4) Compute an estimate of $\mathbf{W}$ by $(\widehat{\mathbf{G}}^T\widehat{\mathbf{A}})^{-1}(\widehat{\mathbf{G}}^T\widehat{\mathbf{G}})$.
(5) Do the LiNGAM permutation and re-normalizing on estimated $\mathbf{W}$ to get an estimate of $\mathbf{B}$:
   (a) Find the one and only permutation of rows of the estimated $\mathbf{W}$ which yields a matrix $\widetilde{\mathbf{W}}$ without any zeros on the main diagonal. In practice, the permutation is sought which minimizes $\sum_i 1/|\widetilde{\mathbf{W}}_{ii}|$.
   (b) Divide each row of $\widetilde{\mathbf{W}}$ by its corresponding diagonal element, to yield a new matrix $\widetilde{\mathbf{W}}'$ with all ones on the diagonal.
   (c) Compute an estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ using $\widehat{\mathbf{B}} = \mathbf{I}-\widetilde{\mathbf{W}}'$.

In the next section, we conduct simulations to see the empirical performance.

6 Simulation experiments

As a sanity check of our method, we performed an experiment with simulated data. We generated data in the following manner:

(1) First, we randomly constructed a strictly lower-triangular matrix $\mathbf{B}$ for four latent factors so that standard deviations of factors f_i owing to parent factors ranged in the interval $[0.5, 1.5]$ and also randomly selected standard deviations of the external influence variables d_i from the interval $[0.5, 1.5]$. Then we normalized the factors so that they were of unit variance. We made the factor correlations range from -0.81 to 0.81. Both fully connected and sparse networks were created.

(2) Next, we generated data with sample size 1,000 by independently drawing the external influence variables d_i from various non-Gaussian distributions with zero mean and unit variance [2] . The values of the latent factors f_i were generated according to the assumed recursive process.

(3) We randomly permuted the order of the factors f_i to hide the causal order with which the data was generated. We also permuted $\mathbf{B}$ as well as the variances of the external influence variables to match the new order.

(4) We generated measurement errors (e_i) in the same manner as generating external influences (d_i) in Step 2. We randomly generated the factor loading matrix $\mathbf{G}$ with pure and impure measurement variables. The number of pure measurement variables was 4 for each latent factor. The number of impure measurement variables that have all the latent factor as their parent was 8. Thus, the total number of observed variables was 24.

(5) We normalized the rows of $\mathbf{G}$ so that variances of x_i were 1 when measurement errors above were added. We varied $\text{var}(e_i)$ from 0.2 to 0.6.

(6) We mixed the generated latent factors f_i and measurement errors e_i to create the observed data x_i.

===== Insert Figure 2 =====

Fig. 2 gives scatterplots of the elements of estimated $\mathbf{B}$ and $\mathbf{G}$ versus the generating ones. The left is the scatterplot of the estimated connection strengths b_{ij} versus the original (generating) values. The right is the scatterplot of the

[2] We first generated a Gaussian variable z with zero mean and unit variance and subsequently transformed it to a non-Gaussian variable by $e_i = \text{sign}(z)|z|^q$. The nonlinear exponent q was selected to lie in $[0.5, 0.8]$ or $[1.2, 2.0]$. The former gave a sub-Gaussian variable, and the latter a super-Gaussian variable. Finally, the transformed variable was standardized to have zero mean and unit variance.

estimated factor loadings g_{ij} versus the generating values. We can see that most of the data points are close enough to the main diagonal, which confirms the validity of our estimation procedure.

7 Conclusion

In this paper, we showed that a linear non-Gaussian acyclic model for latent factors is completely identified. This would be an important step for developing advanced methods to discover structures of latent factors.

References

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

[2] M. Davies, Identifiability issues in noisy ICA, IEEE Signal Processing Letters 11 (5) (2004) 470–473.

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

[4] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Trans. on Neural Networks 10 (1999) 626–634.

[5] A. Hyvärinen, J. Karhunen, E. Oja, Independent component analysis, Wiley, New York, 2001.

[6] A. Mooijaart, Factor analysis for non-normal variables, Psychometrika 50 (1985) 323–342.

[7] E. Moulines, J.-F. Cardoso, E. Gassiat, Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models, in: Proc. ICASSP'97, Munich, Germany, 1997.

[8] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2000.

[9] T. Reilly, R. M. O'Brien, Identification of confirmatory factor analysis models of arbitrary complexity, Sociological Methods & Research 24 (4) (1996) 473–491.

[10] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, A linear non-gaussian acyclic model for causal discovery, J. Mach. Lear. Res. 7 (2006) 2003–2030.

[11] R. Silva, R. Scheines, C. Glymour, P. Spirtes, Learning the structure of linear latent variable models, J. Mach. Lear. Res. 7 (2006) 191–246.

[12] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, Springer Verlag, 1993.

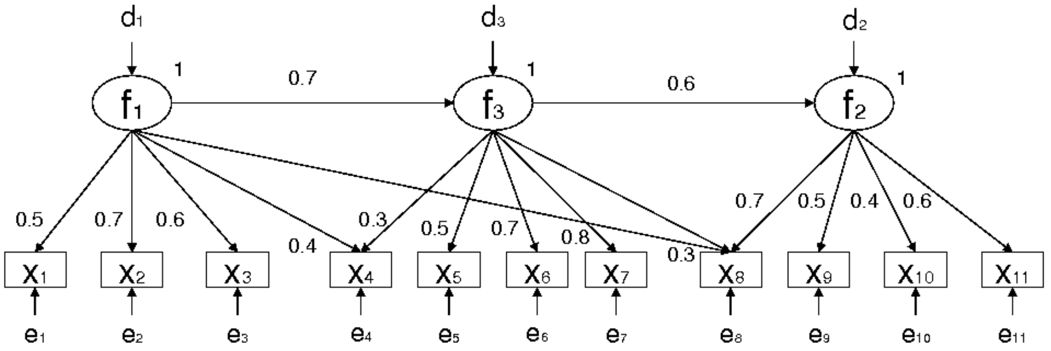

Fig. 1. A graphical example of our data generating models. Latent factors are enclosed in circles, and observed variables are in rectangular boxes. The x_4 and x_8 are impure measurement variables that measure more than one latent factor, and the other observed variables are pure measurement variables that have a single latent factor parent.

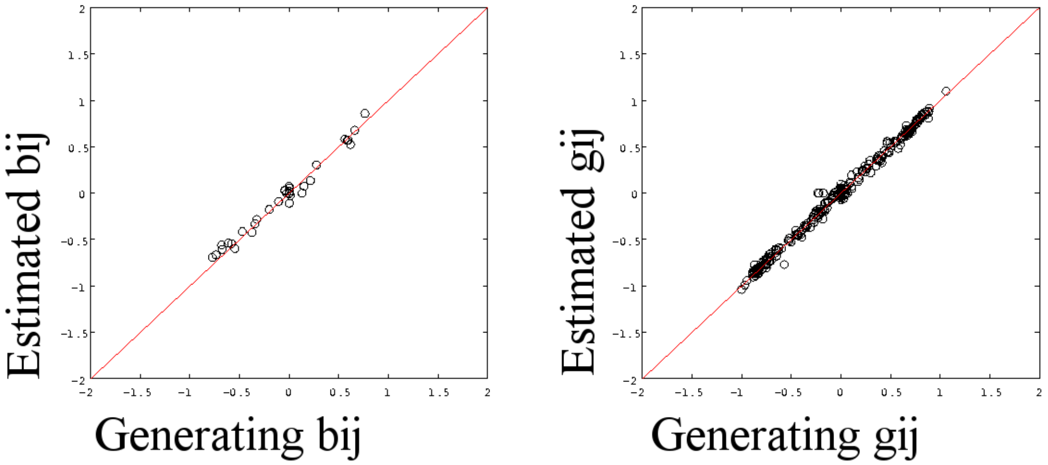

Fig. 2. Left: Scatterplot of the estimated b_{ij} versus the original values. Right: Scatterplot of the estimated g_{ij} versus the original (generating) values. Five data sets with 1,000 observations were generated for each of the scatterplots. Since the ordering and signs of latent factors cannot be defined, we first permuted columns of estimated **G** and multiplied them by -1 if necessary so that the difference between estimated **G** and generating **G** was minimized and compared estimated and generating **B**.