Discovery of linear non-gaussian acyclic models in the presence of latent classes

Shohei Shimizu[1,2], Aapo Hyvärinen[1]

[1] Helsinki Institute for Information Technology, Finland
[2] The Institute of Statistical Mathematics, Japan
http://www.hiit.fi/neuroinf

Abstract. An effective way to examine causality is to conduct an experiment with random assignment. However, in many cases it is impossible or too expensive to perform controlled experiments, and hence one often has to resort to methods for discovering good initial causal models from data which do not come from such controlled experiments. We have recently proposed such a discovery method based on independent component analysis (ICA) called LiNGAM and shown how to completely identify the data generating process under the assumptions of linearity, non-gaussianity, and no latent variables. In this paper, after briefly recapitulating this approach, we extend the framework to cases where latent classes (hidden groups) are present. The model identification can be accomplished using a method based on ICA mixtures. Simulations confirm the validity of the proposed method.

1 Introduction

An effective way to examine causality is to conduct an experiment with random assignment [1]. However, in many cases it is impossible or too expensive to perform controlled experiments. Hence one often has to resort to methods for discovering good initial causal models from data which do not come from such controlled experiments, though obviously one can never fully prove the validity of a causal model from such uncontrolled data alone. Thus, developing methods for causal inference from uncontrolled data is a fundamental problem with a very large number of potential applications such as social sciences [2], gene network estimation [3] and brain connectivity analysis [4].

Previous methods developed for statistical causal analysis of non-experimental data [2, 5, 6] generally work in one of two settings. In the case of discrete data, no functional form for the dependencies is usually assumed. On the other hand, when working with continuous variables, a linear-Gaussian approach is almost invariably taken and has hence been based solely on the covariance structure of the data. Because of this, additional information (such as the time-order of the variables and prior information) is usually required to obtain a full causal model of the variables. Without such information, algorithms based on the Gaussian assumption cannot in most cases distinguish between multiple equally possible causal models.

We have recently shown that when working with continuous-valued data, a significant advantage can be achieved by departing from the Gaussianity assumption [7–9]. The linear-Gaussian approach usually only leads to a *set* of possible models that are equivalent in their covariance structure. The simplest such case is that of two variables, x_1 and x_2. A method based only on the covariance matrix has no way of preferring $x_1 \to x_2$ over the reverse model $x_1 \leftarrow x_2$ [2, 7]. However, a linear-*non-Gaussian* setting actually allows the linear acyclic model to be uniquely identified [9].

In this paper, we extend our previous work to cases where latent classes (hidden groups) are present. The paper is structured as follows. In Section 2 we briefly describe the basics of LiNGAM and subsequently extend the framework in Section 3. Some illustrative examples are provided in Section 4, and the proposed method is empirically evaluated in Section 5. Section 6 concludes the paper.

2 LiNGAM

Here we provide a brief review of our previous work [9]. Assume that we observe data generated from a process with the following properties:

1. The observed variables x_i, $i = \{1 \dots n\}$ can be arranged in a causal order $k(i)$, defined to be an ordering of the variables such that no later variable in the order participates in generating the value of any earlier variable. That is, the generating process is *recursive* [2], meaning it can be represented graphically by a *directed acyclic graph* (DAG) [5, 6].
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' (noise) term e_i, and plus an optional constant term μ_i, that is
$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i + \mu_i. \tag{1}$$
3. The disturbances 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, i.e. $p(e_1, \dots, e_n) = \prod_i p_i(e_i)$.

A model with these three properties we call a *Linear, Non-Gaussian, Acyclic Model*, abbreviated LiNGAM.

We assume that we observe a large number of data vectors $\mathbf{x}$ (containing the components x_i), and each is generated according to the above described process, with the same causal order $k(i)$, same coefficients b_{ij}, same constants μ_i, and the disturbances e_i sampled independently from the same distributions. Note that the above assumptions imply that there are *no unobserved (latent) confounders* [5] (hidden variables). Spirtes et al. [6] call this the *causally sufficient* case.

To see how we can identify the parameters of the model from the set of data vectors $\mathbf{x}$, we start by subtracting out the mean of each variable x_i, leaving us with the following system of equations:

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

where $\mathbf{B}$ is a matrix that contains the coefficients b_{ij} and that could be permuted (by simultaneous equal row and column permutations) to strict lower triangularity if one knew a causal ordering $k(i)$ of the variables. (Strict lower triangularity is here defined as lower triangular with all zeros on the diagonal.) Solving for $\mathbf{x}$ one obtains

$$\mathbf{x} = \mathbf{A}\mathbf{e}, \tag{3}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$. Again, $\mathbf{A}$ could be permuted to lower triangularity (although not *strict* lower triangularity, actually in this case all diagonal elements will be *non-zero*) with an appropriate permutation $k(i)$. Taken together, equation (3) and the independence and non-gaussianity of the components of $\mathbf{e}$ define the standard linear independent component analysis (ICA) model [10,11], which is known to be identifiable.

While ICA is essentially able to estimate $\mathbf{A}$ (and $\mathbf{W} = \mathbf{A}^{-1}$), there are two important indeterminacies that ICA cannot solve: First and foremost, the order of the independent components is in no way defined or fixed. Thus, we could reorder the independent components and, correspondingly, the columns of $\mathbf{A}$ (and rows of $\mathbf{W}$) and get an equivalent ICA model (the same probability density for the data). In most applications of ICA, this indeterminacy is of no significance and can be ignored, but in LiNGAM, we can and we have to find the correct permutation as described in [9]: the correct permutation is the only one which has no zeros in the diagonal.

The second indeterminacy of ICA concerns the scaling of the independent components. In ICA, this is usually handled by assuming all independent components to have unit variance, and scaling $\mathbf{W}$ and $\mathbf{A}$ appropriately. On the other hand, in LiNGAM (as in structural equation modeling, SEM [2]) we allow the disturbance variables to have arbitrary (non-zero) variances, but fix their weight (connection strength) to their corresponding observed variable to unity. This requires us to re-normalize the rows of $\mathbf{W}$ so that all the diagonal elements equal unity in order to obtain $\mathbf{B}$.

Our LiNGAM discovery algorithm [9] can thus be briefly summarized: First, use a standard ICA algorithm to obtain an estimate of the demixing matrix $\mathbf{W}$, permute its rows such that there are no zeros on its diagonal, rescale each row by dividing by the element on the diagonal, and finally compute $\mathbf{B} = \mathbf{I} - \mathbf{W}'$, where $\mathbf{W}'$ denotes the permuted and rescaled $\mathbf{W}$. To find a causal order $k(i)$ we must subsequently find a second permutation, to be applied equally both to the rows and columns of $\mathbf{B}$, which yields strict lower triangularity.

3 LiNGAM in the presence of latent classes

In this section, we extend the basic LiNGAM above to cases where latent (hidden) classes are present.

3.1 Motivation

Let us begin by an example. Regarding a child's and a parent's height, earlier studies (e.g., [12]) pointed out that there is a hereditary effect on height, which is especially stronger between a child and the same-sex parent. This implies that the connection strengths from parent's height to child's height (and possibly the network structures) could be different between the two classes (same-sex and different-sex children). This is a nonlinear relation between child's and parent's height even if the relations are still linear in each class, which cannot be found if the class-membership is ignored (see Section 4 for some artificial examples). In cases where such class-membership is observed, we only have to analyze each class separately. However, in many cases, it would be quite difficult to detect and observe class-membership especially before collecting data. Thus, we need a sophisticated method to estimate latent classes in a data-driven way. In the following, we extend the basic LiNGAM so that the method can estimate latent classes of samples that have similar network structures.

3.2 Model

Let us assume that the data are generated by the following mixture density:

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{B}_k) p(C = k), \tag{4}$$

where $\Theta = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k]$, $\boldsymbol{\theta}_k = [\boldsymbol{\mu}_k^T, \mathrm{vec}(\mathbf{B}_k)^T]^T$, $\boldsymbol{\mu}_k$ is a mean vector, $\mathbf{B}_k$ is a connection strength matrix for class k and C is a discrete variable that indicates the class $k = 1, \cdots, K$. (The $\mathrm{vec}(\cdot)$ denotes the vectorization operator which creates a column vector from a matrix by stacking its columns.) Here, we do *not* assume that we know the number of latent classes K and a priori probability $p(C = k)$. Moreover, the data within class k are assumed to be generated by the LiNGAM model:

$$\mathbf{x} = \mathbf{B}_k \mathbf{x} + (\mathbf{I} - \mathbf{B}_k)\boldsymbol{\mu}_k + \mathbf{e}_k, \tag{5}$$

where $\mathbf{e}_k$ is the disturbance (error) vector for class k. Note that the means, connection strengths and structure of the network ($\boldsymbol{\mu}_k$ and $\mathbf{B}_k$) can be different between classes. See Section 4 for some illustrative examples.

3.3 Model identification using ICA mixtures

We propose that the new model above can be estimated using ICA mixture models [13]. As in the basic LiNGAM, ICA model holds for each class:

$$\mathbf{x} = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{e}_k, \tag{6}$$

where $\mathbf{A}_k = (\mathbf{I} - \mathbf{B}_k)^{-1}$. Then the mixture density is just the ICA mixture model [13]. After $\boldsymbol{\mu}_k$ and $\mathbf{A}_k$ are estimated, we can obtain estimates of $\mathbf{B}_k$

and causal orderings $k(i)$ for class k in the same manner as the basic LiNGAM (Section 2).

Some estimation methods for the ICA mixtures have been proposed [13, 14]. Here we employ the minimum β-divergence method [14] since the β-divergence method does *not* require that the number of classes K and a priori probability $p(C = k)$ are known, which is a big advantage over [13]. Some drawbacks are that one has to tell the algorithm whether the disturbances e_i are super- or sub-gaussian and select a tuning parameter β using a cross-validation technique [15]. Fortunately, the first problem can be solved by (possibly non-parametric) estimation of the source densities [14, 16].

4 Illustrative examples

In this section, we provide two illustrative examples of the LiNGAM in the presence of latent classes (abbreviated as LcLiNGAM) proposed above. We selected $\boldsymbol{\mu}_k$ and $\mathbf{B}_k$ manually as explained below. The disturbances followed the Laplace distribution with zero means and selected the variances so that observed variables had unit variances. Moreover, the number of latent classes was 2, and 250 data points were generated for each class. Note that the numbers of latent classes were estimated as well by the β-divergence method [14]. The scatterplots of observed variables were shown in Figure 1.

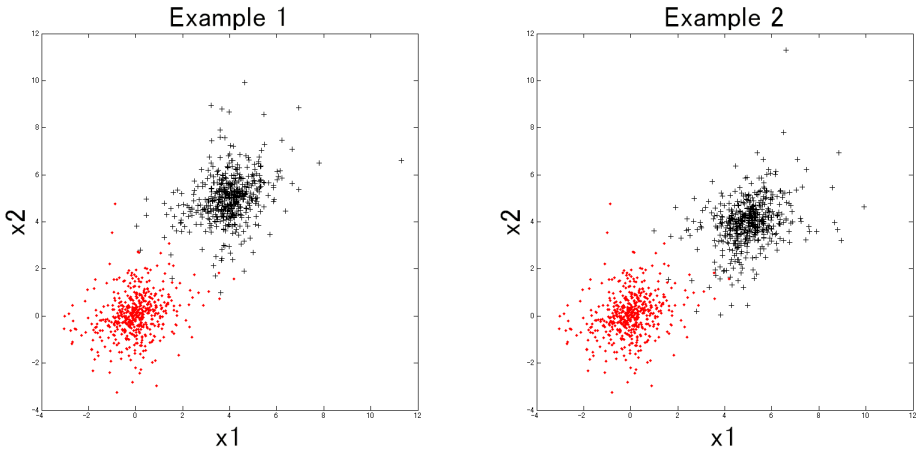

Fig. 1. Left: Scatterplots of the observed variables in Example 1. Right: Scatterplots of the observed variables in Example 2. In the scatterplots, "." denote members of class 1 and "+" those of class 2.

4.1 Example 1

We generated data using the following means, connection strengths and structures of networks:

$$\textbf{Class 1}: \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{B}_1 = \begin{bmatrix} 0 & 0 \\ 0.3 & 0 \end{bmatrix} \tag{7}$$

$$\textbf{Class 2}: \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \ \mathbf{B}_2 = \begin{bmatrix} 0 & 0 \\ 0.3 & 0 \end{bmatrix}. \tag{8}$$

Both classes 1 and 2 had the same causal orders $x_1 \rightarrow x_2$, but different means ($\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$). The different mean structures created a strong correlation (0.88) for the whole data, although the connection strength in each class was rather weak (0.3).

The estimation results by the LcLiNGAM and basic LiNGAM were as follows: LcLiNGAM[3]:

$$\textbf{Class 1}: \quad \boldsymbol{\mu}_1 = \begin{bmatrix} -0.02 \\ 0.06 \end{bmatrix}, \ \mathbf{B}_1 = \begin{bmatrix} 0 & 0 \\ 0.30 & 0 \end{bmatrix} \tag{9}$$

$$\textbf{Class 2}: \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 4.01 \\ 5.01 \end{bmatrix}, \ \mathbf{B}_2 = \begin{bmatrix} 0 & 0 \\ 0.41 & 0 \end{bmatrix}, \tag{10}$$

LiNGAM:

$$\boldsymbol{\mu} = \begin{bmatrix} -0.09 \\ 3.99 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 & 0.81 \\ 0 & 0 \end{bmatrix}. \tag{11}$$

The LcLiNGAM successfully recovered the means and structures of the networks and estimated connection strengths fairly well for both latent classes. However, the basic LiNGAM failed to find the correct causal order and overestimated the connection strength.

4.2 Example 2

Next, we tried data whose means, connection strengths and structures of network were as follows:

$$\textbf{Class 1}: \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \mathbf{B}_1 = \begin{bmatrix} 0 & 0 \\ 0.3 & 0 \end{bmatrix} \tag{12}$$

$$\textbf{Class 2}: \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}, \ \mathbf{B}_2 = \begin{bmatrix} 0 & 0.3 \\ 0 & 0 \end{bmatrix}. \tag{13}$$

Now the two classes had the different causal orders: $x_1 \rightarrow x_2$ for class 1 and $x_1 \leftarrow x_2$ for class 2. The connection strengths were the same but the mean structures were different between the classes.

[3] Obviously, the orders of latent classes are not recovered. In the examples, for the clarity, we permuted the classes so that the differences of estimates and true values were minimized.

The estimation results were as follows:
LcLiNGAM:

$$\textbf{Class 1}: \quad \boldsymbol{\mu}_1 = \begin{bmatrix} -0.02 \\ 0.07 \end{bmatrix}, \ \mathbf{B}_1 = \begin{bmatrix} 0 & 0 \\ 0.39 & 0 \end{bmatrix} \tag{14}$$

$$\textbf{Class 2}: \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 5.01 \\ 4.01 \end{bmatrix}, \ \mathbf{B}_2 = \begin{bmatrix} 0 & 0.41 \\ 0 & 0 \end{bmatrix}, \tag{15}$$

LiNGAM:

$$\boldsymbol{\mu} = \begin{bmatrix} 3.88 \\ 0.08 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 & 0.78 \\ 0 & 0 \end{bmatrix}. \tag{16}$$

The LcLiNGAM estimated the connection strengths fairly well and found correct causal orders for each class. However, the basic LiNGAM could not find that the two classes have different causal orders because it cannot represent any difference between the classes; it estimated, rather arbitrarily, one single causal order $x_1 \leftarrow x_2$.

5 Simulation

To further verify the validity of our method, we performed experiments with simulated data. We repeatedly performed the following experiment:

1. First, we randomly constructed a strictly lower-triangular matrix $\mathbf{B}$ for each class, where the number of classes was 2 and the number of variables was 4. We also randomly selected variances of the disturbance variables. We further generated values for the constants μ_i making the classes have small overlap.[4]
2. Next, we generated data with sample size 500 by independently drawing the disturbance variables e_i from the uniform distribution with zero mean and unit variance for each class. The observed data $\mathbf{X}$ were generated according to the assumed recursive process and were combined to create a whole data.
3. Finally, we fed the data to our discovery algorithm. The β-divergence method was employed to estimate ICA mixtures. Here we told the algorithm that the disturbances were sub-gaussian.
4. We compared the estimated parameters to the generating parameters. In particular, we made a scatterplot of the entries in the estimates $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\mathbf{B}}_k$ against the corresponding ones in $\boldsymbol{\mu}_k$ and $\mathbf{B}_k$. (Note that the numbers of latent classes were estimated as well.)

Figure 2 gives scatterplots of the elements of estimated $\boldsymbol{\mu}_k$ and $\mathbf{B}_k$ versus the generating ones. The left is the scatterplot of the estimated means μ_i versus the original (generating) values. The right is the scatterplot of the estimated connection strengths b_{ij} versus the original (generating) values. We can see that the estimation works well, as evidenced by the grouping of the data points onto the main diagonal.

[4] We first set $\boldsymbol{\mu}_1 = \mathbf{0}$ and took as the elements of $\boldsymbol{\mu}_2$ 1.5 times the sum of standard deviations of corresponding observed variables of each class multiplied by -1 with probability 50%.

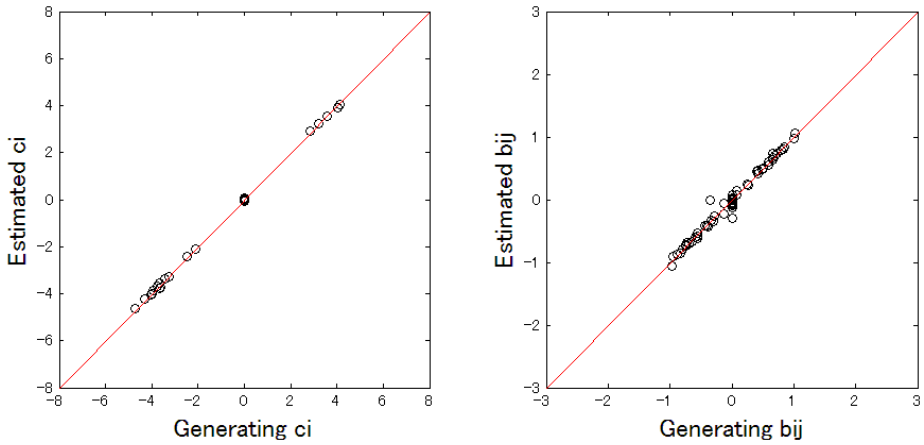

Fig. 2. Left: Scatterplots of the estimated μ_i versus the original (generating) values. Right: Scatterplots of the estimated b_{ij} versus the original (generating) values. Five data sets were generated for the scatterplots.

6 Conclusion

Developing statistical causal inference methods based on non-experimental data is a fundamental problem with a large number of potential applications. Previous methods developed for linear causal models [2, 5, 6] have been based on an explicit or implicit assumption of gaussianity, and have hence been based solely on the covariance structure of the data. Therefore, algorithms based on the gaussian assumption cannot in most cases distinguish between multiple equally possible causal models. In previous work, we have shown that an assumption of non-gaussianity of the disturbance variables, together with the assumption of linearity and no latent variables, allows the linear acyclic model to be completely identified [9].

In this paper, we extended our previous work to cases where latent (hidden) classes are present. The new method can identify the DAG structures within latent classes and would enjoy a wider variety of applications.

Although in the artificial experiments our method worked well, obviously we need to evaluate its empirical performance by more extensive simulations as well as real-world data. For example, in many real situations latent classes would be much more overlapping than in the simulations. Unfortunately, however, for such heavily overlapping cases, ICA estimation methods are still under development [14]. These are important topics for future research.

As a further analysis, it is quite important to investigate what characterizes the latent classes in order to understand how the model can be applied, for example, in the design of practical interventions. The estimated means, connection strengths and structures of networks could provide an interpretation of the latent classes. For example, in Examples 1 and 2 above (Section 4), the differences

of the means would be useful to interpret the difference of the two classes (and probably classify new samples). An additional or alternative way is to analyze the samples classified into the latent classes using logistic regression analysis if some covariates such as sex and age are available. One direction of future research would be to combine the latent class LiNGAM and logistic regression to improve the class distinction ability.[5]

A related topic is the case where hidden confounding (continuous) variables are present (Latent variable LiNGAM) [18]. We would like to mention a useful connection between the two extensions of the basic LiNGAM. In the latent class LiNGAM discussed here, we basically have a binary (discrete) hidden confounding variable (=class membership) which determines the connection strengths when the structure of the network is the same for the different classes. In future work, we will consider a unifying framework that combines the two extensions.

Acknowledgment

This work was partially carried out at Transdisciplinary Research Integration Center, Research Organization of Information and Systems. The authors would like to thank Patrik Hoyer for his valuable comments and Nurul Mollah, Mihoko Minami and Shinto Eguchi for providing access to their Matlab code for ICA mixtures. S.S. was supported by Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science.

References

1. Holland, P.W.: Statistics and causal inference (with discussion). Journal of the American Statistical Association **81** (1986) 945–970
2. Bollen, K.A.: Structural Equations with Latent Variables. John Wiley & Sons (1989)
3. Imoto, S., Goto, T., Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. In: Proc. Pacific Symposium on Biocomputing. Volume 7. (2002) 175–186
4. Kim, J., Zhu, W., Chang, L., Bentler, P.M., Ernst, T.: Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. Human Brain Mapping **28** (2007) 85–93
5. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
6. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd ed. MIT Press (2000)
7. Shimizu, S., Kano, Y.: Use of non-normality in structural equation modeling: Application to direction of causation. Journal of Statistical Planning and Inference (2006) In press.

[5] Such a combination of mixture modeling and logistic regression has been proposed in the context of structural equation modeling (SEM) [17], although such SEM that are based on gaussian mixtures requires that causal orders are prespecified since non-gaussianity is not utilized for the model identification.

8. Shimizu, S., Hyvärinen, A., Hoyer, P.O., Kano, Y.: Finding a causal ordering via independent component analysis. Computational Statistics & Data Analysis **50**(11) (2006) 3278–3293
9. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. J. of Machine Learning Research **7** (2006) 2003–2030
10. Comon, P.: Independent component analysis. a new concept? Signal Processing **36** (1994) 62–83
11. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley, New York (2001)
12. Tanner, J., Israelsohn, W.: Parent-child correlation for body measurements of children between the ages one month and seven years. Ann.Hum.Genet. **26** (1963) 245–259
13. Lee, T.W., Lewicki, M., Sejnowski, T.: ICA mixture models for unsupervised classification of non-gaussian sources and automatic context switching in blind signal separation. IEEE Trans. on Pattern Recognition and Machine Intelligence **22**(10) (2000) 1–12
14. Mollah, M.N.H., Minami, M., Eguchi, S.: Exploring latent structure of mixture ICA models by the minimum β-divergence method. Neural Computation **18** (2006) 166–190
15. Minami, M., Eguchi, S.: Adaptive selection for minimum β-divergence method. In: Proc. the Fourth Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2003). Nara, Japan. (2003) 475–480
16. Pham, D.T., Garrat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. Signal Processing **45** (1997) 1457–1482
17. Muthén, B.O.: Beyond SEM: General latent variables modeling. Behaviormetrika **29** (2002) 81–117
18. Hoyer, P.O., Shimizu, S., Kerminen, A.: Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. In: Proc. the third European Workshop on Probabilistic Graphical Models (PGM2006). (2006) 155–162