

Discovery of Exogenous Variables in Data with More Variables than Observations

Yasuhiro Sogawa¹, Shohei Shimizu¹, Aapo Hyvärinen², Takashi Washio¹,
Teppei Shimamura³, and Seiya Imoto³

¹ The Institute of Scientific and Industrial Research, Osaka University, Japan

² Dept. Comp. Sci. Dept. Math. and Stat., University of Helsinki, Finland

³ Human Genome Center, Institute of Medical Science, University of Tokyo, Japan

Abstract. Many statistical methods have been proposed to estimate causal models in classical situations with fewer variables than observations. However, modern datasets including gene expression data increase the needs of high-dimensional causal modeling in challenging situations with orders of magnitude more variables than observations. In this paper, we propose a method to find exogenous variables in a linear non-Gaussian causal model, which requires much smaller sample sizes than conventional methods and works even when orders of magnitude more variables than observations. Exogenous variables work as triggers that activate causal chains in the model, and their identification leads to more efficient experimental designs and better understanding of the causal mechanism. We present experiments with artificial data and real-world gene expression data to evaluate the method.

Key words: Bayesian networks, independent component analysis, non-Gaussianity, data with more variables than observations

1 Introduction

Many empirical sciences aim to discover and understand causal mechanisms underlying their objective systems such as natural phenomena and human social behavior. An effective way to study causal relationships is to conduct a controlled experiment. However, performing controlled experiments is often ethically impossible or too expensive in many fields including bioinformatics [1] and neuroinformatics [2]. Thus, it is necessary and important to develop methods for causal inference based on the data that do not come from such controlled experiments.

Many methods have been proposed to estimate causal models in classical situations with fewer variables than observations ($p < n$, p : the number of variables and n : the number of observations). A linear acyclic model that is a special case of Bayesian networks is typically used to analyze causal effects between continuous variables [3, 4]. Estimation of the model commonly uses covariance structure of data only and in most cases cannot identify the full structure (edge directions and connection strengths) of the model with no prior knowledge on the

structure [3, 4]. In [5], the authors proposed a non-Gaussian variant of Bayesian networks called LiNGAM and showed that the full structure of a linear acyclic model is identifiable based on non-Gaussianity without pre-specifying any edge directions between the variables, which is a significant advantage over the conventional methods [4, 3].

However, most works in statistical causal inference including Bayesian networks have discussed classical situations with fewer variables than observations ($p < n$), whereas modern datasets including microarray gene expression data increase the needs of high-dimensional causal modeling in challenging situations with orders of magnitude more variables than observations ($p \gg n$) [1, 2]. Here we consider situations in which p is on the order of 1,000 or more, while n is around 50 to 100. For such high-dimensional data, the previous methods are often computationally intractable or statistically unreliable.

In this paper, we propose a method to find exogenous variables in a linear non-Gaussian causal model, which requires much smaller sample sizes than conventional methods and works even when $p \gg n$. The key idea is to identify which variables are exogenous instead of estimating the entire structure of the model. The simpler task of finding exogenous variables than that of the entire model structure would require fewer observations to work reliably. The new method is closely related to a fairly recent statistical technique called independent component analysis (ICA).

Exogenous variables work as triggers that activate a causal chain in the model, and their identification leads to more efficient experimental designs of practical interventions and better understanding of the causal mechanism. A promising application of Bayesian networks for gene expression data is detection of drug-target genes [1]. The new method proposed in this paper can be used to find which genes a drug first affects and how it triggers the gene network.

The paper is structured as follows. We first review ICA and linear causal models in Section 2. We then define a non-Gaussian causal model and propose a new algorithm to find exogenous variables in Section 3. The performance of the algorithm is evaluated by experiments on artificial data and real-world gene expression data in Sections 4 and 5. Section 6 concludes the paper.

2 Background principles

2.1 Independent component analysis

Independent component analysis (ICA) [6] is a statistical technique originally developed in signal processing. ICA model for a p -dimensional observed continuous random vector $\mathbf{x}$ is defined as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad (1)$$

where $\mathbf{s}$ is a p -dimensional continuous random vector whose components s_i are mutually independent and non-Gaussian and are called independent components, and $\mathbf{A}$ is a constant $p \times p$ invertible matrix. Without loss of generality, we

assume s_i to be of zero mean and unit variance. Let $\widetilde{\mathbf{W}} = \mathbf{A}^{-1}$. Then we have $\mathbf{s} = \widetilde{\mathbf{W}}\mathbf{x}$. It is known that the matrix $\widetilde{\mathbf{W}}$ are identifiable up to permutation of the rows [7].

Let $\widehat{\mathbf{s}} = \mathbf{W}\mathbf{x}$. A major estimation principle for $\widetilde{\mathbf{W}}$ is to find such $\mathbf{W}$ that maximizes the sum of non-Gaussianity of estimated independent components $\widehat{s}_i$, which is known to be equivalent to maximize independence between the estimates when the estimates are constrained to be uncorrelated [6]. In [8], the author proposed a class of non-Gaussianity measures:

$$J(\widehat{s}_i) = J_G(\mathbf{w}_i) = [E\{G(\mathbf{w}_i^T \mathbf{x})\} - E\{G(z)\}]^2, \quad (2)$$

where $\mathbf{w}_i^T$ is the i -th row of $\mathbf{W}$ and is constrained so that $E(\widehat{s}_i^2) = E\{(\mathbf{w}_i^T \mathbf{x})^2\} = 1$ due to the aforementioned assumption on unit variance of s_i , G is a nonlinear and non-quadratic function and z is a Gaussian variable with zero mean and unit variance. In practice, the expectations in Eq. (2) are replaced by their sample means. In the rest of the paper, we say that a variable u is more non-Gaussian than a variable v if $J(u) > J(v)$. The author of [8] further proposed an estimation method based on maximization of non-Gaussianity and proved a theorem to show its (local) consistency:

Theorem 1 *Assume that the input data $\mathbf{x}$ follows the ICA model in Eq. (1). Assume that G is a sufficiently smooth even function. Then the set of local maxima of $J_G(\mathbf{w}_i)$ under the constraint $E\{(\mathbf{w}_i^T \mathbf{x})^2\} = 1$ includes the rows of $\widetilde{\mathbf{W}}$ for which the corresponding independent components s_i satisfy the following condition $E\{s_i g(s_i) - g'(s_i)\} [E\{G(s_i)\} - E\{G(z)\}] > 0$, where $g(\cdot)$ is the derivative of $G(\cdot)$, and $g'(\cdot)$ is the derivative of $g(\cdot)$. $\square$*

Note that any independent component s_i satisfying the condition in Theorem 1 is a *local* maximum of $J_G(\mathbf{w})$ but may not correspond to the *global* maximum. Two conjectures are widely made [6], **Conjecture 1**: the assumption in Theorem 1 is true for most reasonable choices of G and distributions of the s_i ; **Conjecture 2**: the global maximum of $J_G(\mathbf{w})$ is one of s_i for most reasonable choices of G and the distributions of s_i . In particular, if $G(s) = s^4$, Conjecture 1 is true for any continuous random variable whose moments exist and kurtosis is non-zero [8], and it can also be proven that there are no spurious optima [9]. Then the global maximum should be one of s_i , *i.e.*, Conjecture 2 is true as well. However, kurtosis often suffers from sensitivity to outliers. Therefore, more robust functions such as $G(s) = -\exp(-s^2/2)$ are widely used [6].

2.2 Linear acyclic causal models

Causal relationships between continuous observed variables x_i ($i = 1, \dots, p$) are typically assumed to be (i) *linear* and (ii) *acyclic* [3, 4]. For simplicity, we assume that the variables x_i are of zero mean. Let $k(i)$ denote such a causal order of x_i that no later variable causes any earlier variable. Then, the linear causal relationship can be expressed as

$$x_i := \sum_{k(j) < k(i)} b_{ij} x_j + e_i, \quad (3)$$

where e_i is an external influence associated with x_i and is of zero mean. (iii) The *faithfulness* [4] is typically assumed. In this context, the faithfulness implies that correlations and partial correlations between variables x_i are entailed by the graph structure, *i.e.*, the zero/non-zero status of b_{ij} , not by special parameter values of b_{ij} . (iv) The external influences e_i are assumed to be independent, which implies there are *no unobserved confounders* [4].

We emphasize that x_i is equal to e_i if it is not influenced by any other observed variable x_j ($j \neq i$) inside the model, *i.e.*, all the b_{ij} ($j \neq i$) are zeros. That is, an external influence e_i is *observed* as x_i . Then the x_i is called an *exogenous observed* variable. Otherwise, e_i is called an *error*. For example, consider the model defined by

$$\begin{aligned} x_1 &= e_1 \\ x_2 &= 1.5x_1 + e_2 \\ x_3 &= 0.8x_1 - 1.3x_2 + e_3. \end{aligned}$$

x_1 is equal to e_1 since it is not influenced by either x_2 or x_3 . x_1 is an exogenous observed variable, and e_2 and e_3 are errors. Note that it is obvious that there *exists at least one exogenous observed variable* $x_i (= e_i)$ due to the acyclicity and no unobserved confounder assumptions.

3 A new method to identify exogenous variables

3.1 A new non-Gaussian linear acyclic causal model

We make two additional assumptions on the distributions of e_i to the model (3) and define a new non-Gaussian linear causal model. Let the observed variables x_i in a p -dimensional vector be $\mathbf{x}$ and external influences e_i in a p -dimensional vector $\mathbf{e}$. Let a $p \times p$ matrix $\mathbf{B}$ consist of the causal effects b_{ij} where the diagonal elements b_{ii} are all zeros. Then the model (3) is written in a matrix form as:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}. \quad (4)$$

Recall that the set of the external influences e_i consist of both exogenous observed variables and errors. To distinguish the exogenous variables and errors, we make the following additional assumptions, **Assumption 1**: External influences that correspond to exogenous observed variables are non-Gaussian; **Assumption 2**: External influences that correspond to errors are non-Gaussian but less non-Gaussian than the exogenous observed variables. That is, *the model (4)=the model (3)+Assumptions 1 and 2*. The first assumption is made to explain why observed data are often considerably non-Gaussian in many fields [6]. The second assumption reflects two facts: i) in statistics, errors have been typically considered to arise as sums of a number of unobserved (non-Gaussian) independent variables, which is why classical methods assume that errors are Gaussian resorting to the central limit theorem; ii) the distinction between Gaussian and

non-Gaussian variables is artificial in practice, though. In reality, many variables are not exactly Gaussian. Therefore, we allow the errors to be strongly non-Gaussian as long as they are less non-Gaussian than exogenous variables.⁴

The distinction between exogenous variables and errors leads to a very simple estimation of exogenous variables proposed in the next subsections.

3.2 Identification of exogenous variables based on non-Gaussianity and uncorrelatedness

We relate the linear non-Gaussian causal model (4) with ICA similarly to [5]. Let us solve the model (4) for $\mathbf{x}$ and then we have an ICA model represented by Eq. (1) as follows

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e} = \mathbf{A}' \mathbf{e}. \quad (5)$$

Note that $\mathbf{I} - \mathbf{B}$ is invertible since it can be permuted to be lower triangular due to the acyclicity assumption if one knew causal orders $k(i)$ [5] and its diagonal elements are all non-zero (unity). In the next section we propose a new algorithm to find exogenous variables $x_i (= e_i)$ using the relation (5). In this section we present two lemmas that ensures the validity of the algorithm.

Lemma 1 *Assume that the input data $\mathbf{x}$ follows the model (4) and that Conjecture 2 (Section 2.1) is true. Let us denote by V_x the set of all the observed variables x_i . Then, the most non-Gaussian observed variable in V_x is exogenous: $J(x_i)$ is maximum in $V_x \Rightarrow x_i = e_i$. $\square$*

Proof Eq. (5) shows that the model (4) is an ICA model, where external influences e_i are independent components (ICs). The set of the external influences consist of exogenous observed variables and errors. Due to the model assumption (Assumption 2 in Section 3.1), exogenous observed variables are more non-Gaussian than errors. Therefore, the most non-Gaussian *exogenous* observed variable is the most non-Gaussian IC. Next, according to Conjecture 2 that is here assumed to be true, the most non-Gaussian IC, *i.e.*, the most non-Gaussian *exogenous* observed variable, is the global maximum of the non-Gaussianity measure $J(\mathbf{w}^T \mathbf{x}) = J_G(\mathbf{w})$ among such linear combinations of observed variables $\mathbf{w}^T \mathbf{x}$ with the constraint $E\{(\mathbf{w}^T \mathbf{x})^2\} = 1$, which include all the observed variables x_i in V_x . Therefore, the most non-Gaussian observed variable is the most non-Gaussian *exogenous* variable. $\blacksquare$

Lemma 2 *Assume the assumptions of Lemma 1. Let us denote by E a strict subset of exogenous observed variables so that it does not contain at least one exogenous variable. Let us denote by U_E the set of observed variables uncorrelated with any variable in E . Then the most non-Gaussian observed variable in U_E is exogenous: $J(x_i)$ is maximum in $U_E \Rightarrow x_i = e_i$. $\square$*

⁴ It would be rather easy to show that our algorithm in Section 3.3 allows Gaussian errors as well.

Proof First, the set V_x is the union of three disjoint sets: E , U_E and C_E , where C_E is the set of observed variables in $V_x \setminus E$ correlated with a variable in E . By definition, any variable in U_E are not correlated with any variable in E . Since the faithfulness is assumed, the zero correlations are only due to the graph structure. Therefore, there is no directed path from any variable in E to any variable in U_E . Similarly, there is a directed path from each (exogenous) variable in E to a variable in C_E . Next, there can be no directed path from any variable in C_E to any variable in U_E . Otherwise, there would be a directed path from such a variable in E from which there is a directed path to a variable in C_E to a variable in U_E through the variable in C_E . Then, due to the faithfulness, the variable in E must correlate with the variable in U_E , which contradicts the definition of U_E .

To sum up, there is no directed path from any variable in $E \cup C_E$ to any variable in U_E . Since any directed path from the external influence e_i associated with any variable x_i in V_x must go through x_i , there is no directed path from the external influence associated with any variable in $E \cup C_E$ to any variable in U_E . In other words, there can be directed paths from *only* the external influences associated with any variables in U_E to some variables in U_E . Then, we again have an ICA model: $\tilde{\mathbf{x}} = \tilde{\mathbf{A}}' \tilde{\mathbf{e}}$, where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{e}}$ are vectors whose elements are the variables in U_E and corresponding external influences in $\mathbf{e}$ in Eq. (5), and $\tilde{\mathbf{A}}'$ is the corresponding submatrix of $\mathbf{A}'$ in Eq. (5). Recursively applying Lemma 1 shows that the most non-Gaussian variable in U_E is exogenous. ■

To find uncorrelated variables, we simply use the ordinary Gaussianity-based testing method [10] and control the false discovery rate [11] to 5% for multiplicity of tests. Though non-parametric methods [10] is desirable for more rigorous testing in the non-Gaussian setting, we used the Gaussian method that is more computationally efficient and seems to work relatively well in our simulations. Future work would address what is the better testing procedure taking non-Gaussianity into account.

3.3 Exogenous generating variable finder: EggFinder

Based on the discussions in the previous subsection, we propose an algorithm to find exogenous variables one by one, which we call EggFinder (ExoGenous Generating variable Finder):

1. Given V_x , initialize $E = \emptyset$, $U_E^{(1)} = V_x$, and $m := 1$.
2. Repeat until no variables x_i are uncorrelated with exogenous variable candidates, *i.e.*, $U_E^{(m)} = \emptyset$:
 - (a) Find the most non-Gaussian variable x_m in $U_E^{(m)}$:

$$x_m = \arg \max_{x \in U_E^{(m)}} J(x), \quad (6)$$

where J is the non-Gaussianity measure in Eq. (2) with

$$G(x) = -\exp(-x^2/2). \quad (7)$$

- (b) Add the most non-Gaussian variable x_m to E , that is, $E = E \cup \{x_m\}$.
- (c) Let $U_E^{(m+1)}$ to be the set of variables x_i uncorrelated with any variable in E , and $m := m+1$.

In Step 2c, we use the Gaussianity-based testing method and control the false discovery rate to 5%.

4 Experiments on artificial data

We studied the performance of EggFinder when $p \gg n$ under a linear non-Gaussian acyclic model having a sparse graph structure and various degrees of error non-Gaussianity. Many real-world networks such as gene networks are often considered to have scale-free graph structures. However, as far as we know, there is no standard way to create a *directed* scale-free graph. Therefore, we first randomly created a (conventional) sparse directed acyclic graph with $p=1,000$ variables using a standard software Tetrad (<http://www.phil.cmu.edu/projects/tetrad/>). The resulting graph contained 1,000 edges and $\ell=171$ exogenous variables. We randomly determined each element of the matrix $\mathbf{B}$ in the model (4) to follow this graph structure and make the standard deviations of x_i owing to parent observed variables ranged in the interval $[0.5, 1.5]$.

We generated non-Gaussian exogenous variables and errors as follows. We randomly generated a non-Gaussian exogenous observed variable $x_i (=e_i)$ that was sub- or super-Gaussian with probability 50%. We first generated a Gaussian variable z_i with zero mean and unit variance and subsequently transformed it to a non-Gaussian variable by $s_i = \text{sign}(z_i)|z_i|^{q_i}$. The nonlinear exponent q_i was randomly selected to lie in $[0.5, 0.8]$ or $[1.2, 2.0]$ with probability 50%. The former gave a sub-Gaussian symmetric variable, and the latter a super-Gaussian symmetric variable. Finally, the transformed variable s_i was scaled to the standard deviation randomly selected in the interval $[0.5, 1.5]$ and was taken as an exogenous variable. Next, for each error e_i , we randomly generated h ($h=1, 3, 5$ and 50) non-Gaussian variables having unit variance in the same manner as for exogenous variables and subsequently took the sum of them. We then scaled the sum to the standard deviation selected similarly to the cases of exogenous variables and finally took it as an error e_i . A larger h (the number of non-Gaussian variables summed) would generate a less non-Gaussian error due to the central limit theorem.

Finally, we randomly generated 1,000 datasets under each combination of h and n ($n=30, 60, 100$ and 200) and fed the datasets to EggFinder. For each combination, we computed percentages of datasets where all the top m estimated variables were actually exogenous. In Fig. 1, the relations between the percentage and m are plotted for some representative conditions due to the limited space. First, in all the conditions the percentages monotonically decrease when m increases. Second, the percentages generally increase when the sample size n increases. Similar changes of the percentages are observed when the errors are less non-Gaussian. This is reasonable since a larger n enables more accurate

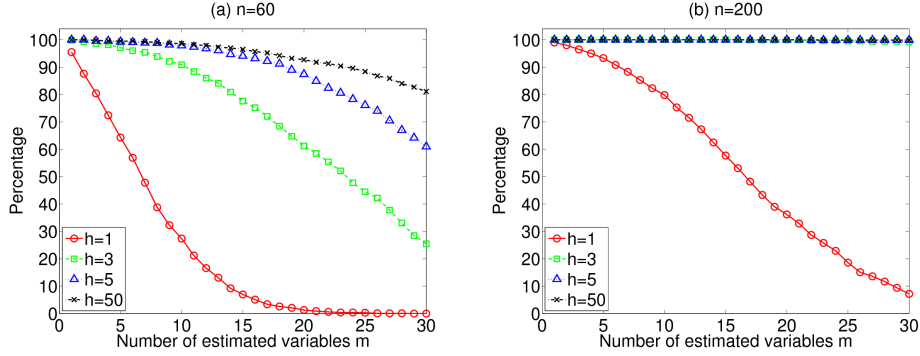


Fig. 1. Percentages of datasets where all the top m estimated variables were actually exogenous under (a) $n=60$; (b) $n=200$.

estimation of non-Gaussianity and correlation, and a larger h generates data more consistent with the assumptions of the model (4). In summary, EggFinder successfully finds a set of exogenous variables up to more than $m=10$ in many practical conditions. However, EggFinder may not find all the exogenous variables when $p \gg n$, although it asymptotically finds all the exogenous variables if all the assumptions made in Lemmas 1 and 2 hold.

Interestingly, EggFinder did not fail completely and identified a couple of exogenous variables even for the $h=1$ condition where the distributional assumption on errors was most likely to be violated. This is presumably because the endogenous variables are sums of non-Gaussian errors and exogenous variables, so due to the central limit theorem they are likely to be less non-Gaussian than the exogenous variables, even if the errors and exogenous variables have the same degree of non-Gaussianity.

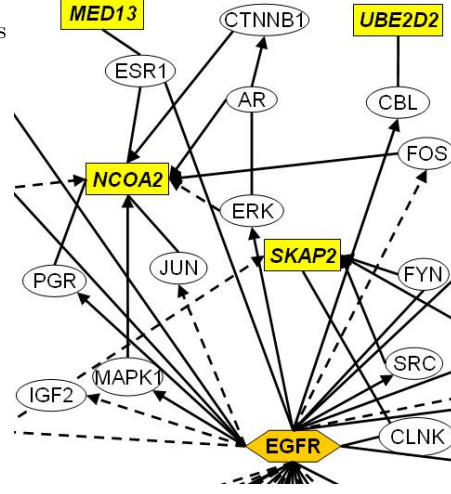
5 Application to microarray gene expression data

To evaluate the practicality of EggFinder, we analyzed a real-world dataset of DNA microarray collected in experiments on human breast cancer cells [12], where epidermal growth factor EGF was dosed to the breast cancer cells, and their gene expression levels were measured. The experiment was conducted with completely random sampling of the cells under every combination of two factors. The first factor was the concentration of EGF (0.1, 0.5, 1.0, and 10.0 nmol/ ℓ), and the second factor was the elapsed time after its dose (5, 10, 15, 30, 45, 60 and 90 minutes). The total number of experimental conditions was 27. No experiment under the condition of the concentration of EGF 10.0 nmol/ ℓ at 45 minutes elapsed time was conducted. For each condition, gene expression levels of 22,277 genes were measured using Affymetrix GeneChip microarrays.

As a standard preprocessing, we first conducted t -tests for the differences of means of the gene expression levels between the lowest and highest concentration

Table 1. Candidates of exogenous genes found by EggFinder.

The genes likely to be exogenous	The others
<i>ACBD3</i>	CAPRIN2
<i>ARPC2</i>	CDC2L6
<i>EIF3M</i>	FKBP15
<i>GULP1</i>	IFT52
<i>MED13</i>	KDM6B
<i>MUT</i>	LOC100134401
<i>NCOA2</i>	LOC202181
<i>NOLC1</i>	PHF20L1
<i>PPIB</i>	PMS2L2
<i>RBMS1</i>	PPDPF
<i>RRM1</i>	PPIH
<i>RSRC1</i>	PPPDE1
<i>SET</i>	RAB14
<i>SKAP2</i>	SH3YL1
<i>UBE2D2</i>	

**Fig. 2.** A part of the pathway network from EGFR to candidates found by EggFinder. The genes boxed and indicated in italic type are the candidates.

conditions of EGF under 5, 10, 15 and 30 minutes elapsed time. We then selected 1,000 genes that expressed the most significance of the differences since such genes were likely to be relevant to EGF dosing. Thus, we obtained a data matrix with the number of variables $p=1,000$ and the sample size $n=27$.

Subsequently, we applied EggFinder to the data matrix. Table 1 shows 29 candidates of exogenous genes found by EggFinder. To evaluate the candidates, we obtained gene pathways from EGF receptor EGFR to the candidates by Ingenuity Pathways Database (<http://www.ingenuity.com/>) which is a literature-based biological pathway database. A part of the gene pathways are shown in Fig. 2 where both a dashed line and a solid line stand for a direct influence from a gene to another gene. A dashed line goes through some intermediate factor such as enzymes, while a solid line does not. In the obtained gene pathway network, 15 of the 29 candidates listed in the left column in Table 1 are reached from EGFR within two edges. These 15 candidates are likely to be exogenous under the biological knowledge. However, it does not mean that the other 14 candidates listed in the right column in Table 1 are not exogenous at all since the biological knowledge on the exogeneity of genes has not been sufficiently accumulated in the database. We merely obtained no strong evidence that the 14 candidates are exogenous by Ingenuity Pathways Database. For instance, among the 14 candidates, CAPRIN2 might be also expected to be exogenous since it is known to be induced by FGF (Fibroblast Growth Factor) similar to EGF [13]. In biological aspects, the relation between EGFR and these 14 candidates are

worth to be examined. By using EggFinder, we can narrow down to the genes worth for examining.

6 Conclusion

We proposed a method to find exogenous variables from data having orders of magnitude more variables than observations. Experiments on microarray gene expression data showed that our method is promising. This would be an important first step for developing advanced causal analysis methods in the challenging situations $p \gg n$.

Acknowledgments. This work was supported in part by the Grant-in-Aid (21700302, 21650029) from the Ministry of Education, Culture, Sports, Science and Technology.

References

1. di Bernardo, D., Thompson, M., Gardner, T., Chobot, S., Eastwood, E., Wojtovich, A., Elliot, S., Schaus, S., Collins, J.: Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotech.* **23** (2005) 377–383
2. Londei, A., D’Ausilio, A., Basso, D., Belardinelli, M.O.: A new method for detecting causality in fMRI data of cognitive processing. *Cog. Proc.* **7** (2006) 42–52
3. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Camb. Univ. Press (2000)
4. Spirtes, P., Glymour, C., Scheines, R.: *Causation, Prediction, and Search*. Springer Verlag (1993)
5. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** (2006) 2003–2030
6. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent component analysis*. Wiley, New York (2001)
7. Comon, P.: Independent component analysis, a new concept? *Signal Processing* **36** (1994) 62–83
8. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks* **10** (1999) 626–634
9. Delfosse, N., Loubaton, P.: Adaptive blind separation of independent sources: a deflation approach. *Signal Processing* **45** (1995) 59–83
10. Lehmann, E., Romano, J.: *Testing Statistical Hypotheses*. Springer (2005)
11. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57** (1995) 289–300
12. Ivshina, A.V., George, J., Senko, O., Mow, B., Putti, T.C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J.E.L., Liu, E.T., Bergh, J., Kuznetsov, V.A., Miller, L.D.: Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* **66** (2006) 10292–10301
13. Lorén, C., Schrader, J., Ahlgren, U., Gunhaga, L.: FGF signals induce Caprin2 expression in the vertebrate lens. *Differentiation* **77** (2009) 386–394