Estimating Exogenous Variables in Data with More Variables than Observations

Yasuhiro Sogawa[a,*], Shohei Shimizu[a], Aapo Hyvärinen[b], Takashi Washio[a], Teppei Shimamura[c], Seiya Imoto[c]

[a] *The Institute of Scientific and Industrial Research, Osaka University, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan*
[b] *Dept. Comp. Sci. Dept. Math. and Stat., University of Helsinki, P.O. Box 68, FIN-00014, Finland*
[c] *Human Genome Center, Institute of Medical Science, University of Tokyo, Shirokanedai 4-6-1, Minato-ku, Tokyo 108-8639, 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 under 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.

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

*Corresponding author. Tel., fax: +81 6 6879 8544
Email address: sogawa@ar.sanken.osaka-u.ac.jp (Yasuhiro Sogawa)

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 (di Bernardo et al., 2005) and neuroinformatics (Londei et al., 2006). Thus, it is 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 (Pearl, 2000; Spirtes et al., 1993). 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 (Pearl, 2000; Spirtes et al., 1993). Shimizu et al. (2006) 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 any prior knowledge, which is a significant advantage over the conventional methods (Pearl, 2000; Spirtes et al., 1993).

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 ($p \gg n$)(di Bernardo et al., 2005; Londei et al., 2006). Here we consider situations in which p is in 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 variables which 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 uses some non-Gaussianity measures developed in 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 (di Bernardo et al., 2005). The new method proposed in this paper can be used to find genes firstly affected by a drug and triggering the gene network.

The paper is structured as follows. We first review some studies on non-Gaussianity 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 using artificial data and real-world gene expression data in Sections 4 and 5. Section 6 concludes the paper.

2. Background principles

2.1. Non-Gaussianity and Negentropy

Probability distributions excluding Gaussian distributions are called non-Gaussian distributions. Any variable which follows a non-Gaussian distribution is called a non-Gaussian variable. Characteristics of the non-Gaussian distributions and the non-Gaussian variables have been extensively studied in the research field of Independent component analysis (ICA) (Hyvärinen et al., 2001). Hyvärinen (1999) proposed a class of non-Gaussianity measures named negentropy to evaluate the non-Gaussian degree of the distribution of a variable $\boldsymbol{x}$:

$$J(\boldsymbol{x}) = [E\{G(\boldsymbol{x})\} - E\{G(z)\}]^2, \tag{1}$$

where G is a nonlinear and non-quadratic function and z is a Gaussian variable with zero mean and unit variance. In practice, $G(s)=-\exp(-s^2/2)$ are widely used for G (Hyvärinen et al., 2001), and the expectations in Eq. (1) 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 $J(u)>J(v)$.*

2.2. Linear acyclic causal models

Causal relationships between continuous observed variables x_i ($i = 1, \cdots, p$) are typically assumed to be (i) *linear* and (ii) *acyclic* (Pearl, 2000; Spirtes

et al., 1993). 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

$$x_i := \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \tag{2}$$

where e_i is an external influence associated with x_i and is of zero mean. (iii) The *'faithfulness'* (Spirtes et al., 1993) is typically assumed. In this context, the faithfulness implies that correlations between variables x_i are entailed by the graph structure, *i.e.*, the zero/non-zero status of b_{ij}. (iv) The external influences e_i are assumed to be independent, which implies there are *'no unobserved confounders'* (Spirtes et al., 1993).

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. Thus, x_1 is an exogenous observed variable, and e_2 and e_3 are errors. Note 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 an additional assumption on the distributions of e_i to the model (2) and define a new non-Gaussian linear causal model. Let a p-dimensional vector x be a set of the observed variables x_i and a p-dimensional vector $\boldsymbol{e}$ be a set of external influences e_i. Let each element b_{ij} of a $p \times p$ matrix $\mathbf{B}$ represent a causal effect from x_j to x_i where the diagonal elements b_{ii} are all zeros. Then the model (2) is written in a matrix form as:

$$\boldsymbol{x} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}. \tag{3}$$

Recall that the set of the external influences e_i consist of both exogenous observed variables and errors. We make the following additional assumption to characterize difference between exogenous variables and errors, **Assumption 1**: External influences that correspond to errors are less non-Gaussian than exogenous variables. That is, *the model (3)=the model (2)+Assumption 1.* The assumption reflects a fact that observed variables are often considerably non-Gaussian in many fields (Hyvärinen et al., 2001). In particular, external influences that correspond to exogenous variables are more non-Gaussian than those corresponding to errors since the 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. However, in reality, almost all the errors are not exactly Gaussian. Therefore, we allow the errors to be non-Gaussian as long as they are less non-Gaussian than exogenous variables as well as to be Gaussian. We further discuss the validity of Assumption 1 in the next subsection. The distinction between exogenous variables and errors leads to a simple estimation of exogenous variables proposed in Section 3.3 and 3.4.

3.2. Central limit theorem for independent and not identically distributed random variables

The assumption in the previous section stating that errors are less non-Gaussian than exogenous variables is supported by a generic nature of the central limit theorem explained in this section. Moreover, non-exogenous observed variables, *i.e.*, endogenous observed variables, are expected to be less non-Gaussian than exogenous variables by the nature of the central limit theorem. A key of these considerations is a condition where the central limit theorem holds. The classical central limit theorem states that the probability distribution of the sum of a sufficiently large number of independent and identically distributed random variables will be approximately Gaussian. However, the identicality among the distributions does not always hold in many practical cases, and thus Gaussianity of the summed variables are not obviously ensured by the theorem. A past study assessed a wider condition called Lindeberg's condition where the sum of such random variables will be Gaussian (Billingsley, 1986). Let us assume that x_k $(k = 1, \cdots, n)$ are independent random variables following its own distribution function F_k which has a finite mean $\mu_k = E[x_k]$ and a finite variance $\sigma_k^2 = \mathrm{Var}[x_k]$. We

denote the sum of the variances by $D_n = \sum_{k=1}^{n} \sigma_k^2$. The Lindberg's condition is as follow.

Theorem 1 (Lindeberg's condition). *If the random variables satisfy the Lindeberg's condition:*

$$\lim_{n\to\infty} \frac{1}{D_n} \sum_{k=1}^{n} \int_{|x_k-\mu_k|\geq\epsilon\sqrt{D_n}} (x_k - \mu_k)^2 dF_k = 0 \quad for \quad \forall\epsilon,$$

the sum of a sufficient number of independent random variables will be Gaussian as $n \to \infty$. □

Though this is only a sufficient condition, the inverse is also true if the random variables x_k satisfies the following condition:

$$\lim_{n\to\infty} \max_{k=1,\ldots,n} \frac{\sigma_k^2}{D_n^2} = 0. \tag{4}$$

That is, the Lindeberg's condition is sufficient and neccesary unless no random variable has a quite large variance nearly equal to an infinite variance. It is expected that the random variables hardly have distributions other than ones having Lindeberg's condition in most cases. Therefore, an error which is a sum of many unobserved independent variables widely tends to be less Gaussian than exogenous observed variables which reflects a unique or a few unobserved (non-Gaussian) independent variables. This supports the aforementioned **Assumption 1** which states that an error which is considered as a sum of unobserved random variables is less non-Gaussian than exogenous variables.

3.3. Identification of exogenous variables based on non-Gaussianity and uncorrelatedness

In this section, we present two lemmas that ensure the validity of our algorithm to find exogenous variables $x_i(= e_i)$. Before showing the lemmans, we introduce a conjecture widely made in the region of Independent Component Analysis (ICA) (Hyvärinen et al., 2001) since the non-Gaussian linear acyclic model proposed in Section2.2 is considered as a variant model of ICA:

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{e}, \qquad (\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}) \tag{5}$$

Let $\widetilde{\mathbf{W}}$=$\mathbf{A}^{-1}$. Then we have $\boldsymbol{e}$=$\widetilde{\mathbf{W}}\boldsymbol{x}$. ICA aims to obtain a demixing matrix $\widetilde{\mathbf{W}}$ from observed data of $\boldsymbol{x}$. A major estimation principle for $\widetilde{\mathbf{W}}$ is to

find such $\mathbf{W}$ that maximizes the sum of non-Gaussianity of external influences $J(\boldsymbol{w}^T\boldsymbol{x})$, which is known to be equivalent to maximize independence between the estimates when the estimates are constrained to be uncorrelated (Hyvärinen et al., 2001). In the region of ICA, a following conjecture is widely made, **Conjecture 1**: If a function G in the non-Gaussianity measure $J(x)$ is a sufficiently smooth even function such as $G(x) = \exp(-x^2/2)$, $\exists e_i \in \boldsymbol{e}$ e_i=arg max$_{\boldsymbol{w}\in\mathbb{R}^p}$ $J(\boldsymbol{w}^T\boldsymbol{x})$. This conjecture states that a sum of even two observed variables is less non-Gaussian than the most non-Gaussian external influence in $\boldsymbol{e}$. Then, we employ this conjecture and show the lemmas and their proof.

Lemma 1. *Let us denote by V_x the set of all the observed variables in the model (3). Assume that the input data of $\boldsymbol{x}$ follows the model (3) and that* **Conjecture 1** *is true. Then, the most non-Gaussian observed variable x_i in V_x is exogenous: $J(x_i)$ is maximum in $V_x \Rightarrow x_i = e_i$.* □

Proof. Due to the model assumption, $\mathbf{B}$ is strictly lower triangular in the model (3). Therefore, there is at least one relation such as $x_i = e_i$. This implies that there is at least one exogenous observed variable in V_x. Because of **Conjecture 1**, $J(e_i) \geq J(\boldsymbol{w}^T\boldsymbol{x})$ for all $\boldsymbol{w} \in \mathbb{R}^p$. This implies $J(e_i) \geq J(x_j)$ for all $x_j \in \boldsymbol{x}$. Since **Assumption 1** states that external influences that correspond to errors are less non-Gaussian than exogenous variables, the most non-Gaussian external influence is an exogenous variable, and therefore, the most non-Gaussian observed variable x_i is an exogenous observed variable e_i, *i.e.*, $J(x_i)$ is maximum in $V_x \Rightarrow x_i = e_i$. ■

Lemma 2. *Assume the assumptions of Lemma 1. Let us denote by E a strict subset of exogenous observed variables in V_x 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 $x_i \in U_E$ is exogenous: $J(x_i)$ is maximum in $U_E \Rightarrow x_i$=e_i.* □

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 $V_x \backslash E$ 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 to a variable in U_E through a variable in C_E. Then, because of 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 our new non-Gaussian linear acyclic causal model: $\widetilde{\boldsymbol{x}}=\widetilde{\mathbf{B}}\widetilde{\boldsymbol{x}} + \widetilde{\boldsymbol{e}}$, where $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{e}}$ are vectors whose elements are the variables in U_E and corresponding external influences in $\boldsymbol{e}$ of Eq. (3), and $\widetilde{\mathbf{B}}$ is the corresponding submatrix of $\mathbf{B}$ in Eq. (3). Recursively applying Lemma 1 shows that the most non-Gaussian variable in U_E is always exogenous. ■

To find uncorrelated variables, we simply use the ordinary Gaussianity-based testing method (Lehmann & Romano, 2005) and control the false discovery rate (Benjamini & Hochberg, 1995) to 5% for multiplicity of tests. Though non-parametric methods (Lehmann & Romano, 2005) 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.

3.4. ExoGenous Generating variable Finder: EggFinder

Based on the discussions in the previous subsection, we propose an algorithm to successively find exogenous observed variables, 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 $U_E^{(m)}$:

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

where J is the non-Gaussianity measure in Eq. (1) with

$$G(x) = -\exp(-x^2/2). \tag{7}$$

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

In Step 2(c), 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 sparse directed acyclic graph with $p{=}1{,}000$ variables using a software Tetrad (http://www.phil.cmu.edu/projects/tetrad/, Accessed in Nov. 14). 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 (3) 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 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 $s_i = \mathrm{sign}(z_i)|z_i|^{q_i}$. 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 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 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 estimation of non-Gaussianity and correlation and generates data more consistent with the assumptions of the model (3). In summary, EggFinder successfully finds a set of exogenous variables up to more than m=10 in many 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 errors and the exogenous variables might satisfy the condition mentioned in Section 3.2, and therefore, endogenous variables, especially the variables being lower in the network, which are sums of errors and exogenous variables are likely to be less non-Gaussian than the exogenous variables due to the central limit theorem, even if the errors and the 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 (Ivshina et al., 2006), 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

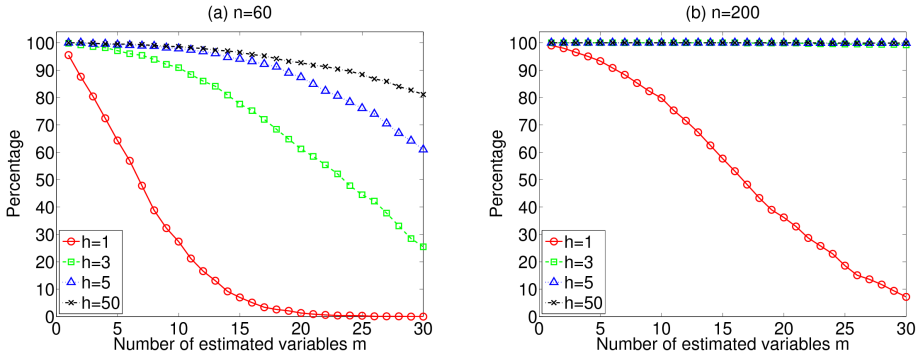

Figure 1: Percentages of datasets where all the top m estimated variables were actually exogenous under (a) n=60; (b) n=200.

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 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 p=1,000 and 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 (IPD, http://www.ingenuity.com/, Accessed in Sep. 30) which is a literature-based biological pathway database. The gene pathway network from EGFR to the candidates is 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 dose 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

Table 1: Candidates of exogenous genes found from the dataset of EGF dosing.

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

Table 2: Candidates of exogenous genes found from the dataset of HRG dosing.

The genes likely to be exogenous	The others
RELA	ARGEF10
CFLAR	HTATSF1
PRPF6	KDM4A
BRCA2	C19ORF40
BUB1	TOR1AIP1
ATF6	CLPTM1
TGOLN2	HNRNPUL2
MAMLD1	N4BP2L2
SYNJ2	PRPF38B
NFAT5	CAMSAP1L1
EIF4B	ZCCHC8
CLIC4	RCLRE1C
EWSR1	POLG2
PSME4	
SART3	
TCF3	
USO1	

evidence that the 14 candidates are exogenous by IPD. 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 (Lorén et al., 2009). In biological aspects, the relation between EGFR and these 14 candidates are worth to be examined.

After evaluating the practicality of EggFinder, we analyzed another dataset of DNA microarray collected in experiments where HRG was dosed to the breast cancer cells instead of EGF. The experiments were conducted in the same manner as for those where EGF was dosed. Unlike the dataset of EGF dosing, there was no lack of experiment, and thus the total number of experimental conditions was 28. We conducted the same preprocessing and selected 1000 genes from 22277 genes that expressed the most significance of the differences of means of the gene expression levels between the lowest and highest concentration conditions of HRG. Then, we obtained a data matrix

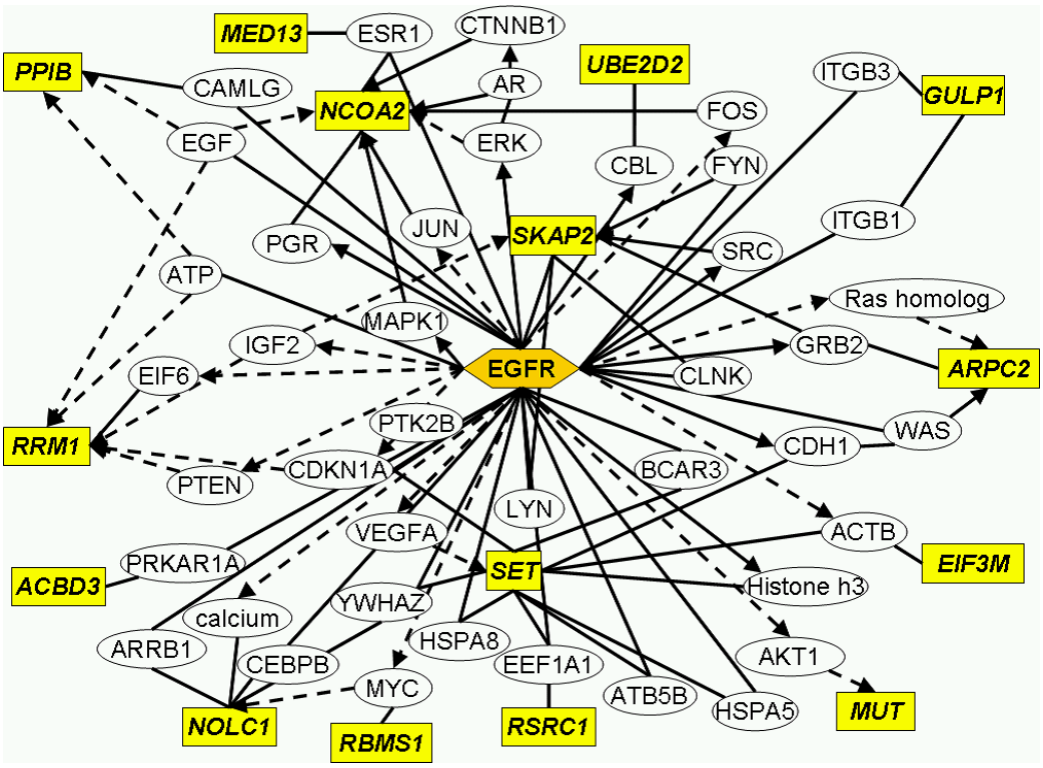

Figure 2: The pathway network from EGFR to candidates found by EggFinder from the dataset of EGF dosing. The genes boxed and indicated in italic type are the candidates.

with the number of variables p=1, 000 and n=28. Following after the analysis for the dataset of EGF dosing, we applied EggFinder to the data matrix to derive up to 30 candidates of exogenous genes.

Eventually, we derived 30 candidates of exogenous genes shown in Table 2. Then, we evaluated the candidates in the same way of the evaluation of ones derived from the dataset of EGF dosing. As a result, 17 of the 30 candidates listed in the left column in Table 2 are likely to be actually exogenous, and the other 13 candidates listed in the right column are worth to be examined. In these manners, we can narrow down to the genes worth for examining by using EggFinder.

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 $p \gg n$.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, *57*, 289–300.

di Bernardo, D., Thompson, M., Gardner, T., Chobot, S., Eastwood, E., Wojtovich, A., Elliot, S., Schaus, S., & Collins, J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotech.*, *23*, 377–383.

Billingsley, P. (1986). *Probability and measure*. Wiley Series in Prob. and Math. Stat.: Probability and Mathematical Statistics.

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

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

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. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.*, *66*, 10292–10301.

Lehmann, E., & Romano, J. (2005). *Testing Statistical Hypotheses*. Springer.

Londei, A., D'Ausilio, A., Basso, D., & Belardinelli, M. O. (2006). A new method for detecting causality in fMRI data of cognitive processing. *Cog. Proc.*, *7*, 42–52.

Lorén, C., Schrader, J., Ahlgren, U., & Gunhaga, L. (2009). FGF signals induce Caprin2 expression in the vertebrate lens. *Differentiation*, *77*, 386–394.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Camb. Univ. Press.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Springer Verlag.