Distinguishing Causes from Effects using Nonlinear Acyclic Causal Models

Kun Zhang KUN.ZHANG@CS.HELSINKI.FI
Dept of Computer Science and HIIT
University of Helsinki
00014 Helsinki, Finland

Aapo Hyvärinen AAPO.HYVARINEN@CS.HELSINKI.FI
Dept of Computer Science, HIIT, and Dept of Mathematics and Statistics
University of Helsinki
00014 Helsinki, Finland

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Distinguishing causes from effects is an important problem in many areas. In this paper, we propose a very general but well defined nonlinear acyclic causal model, namely, post-nonlinear acyclic causal model with inner additive noise, to tackle this problem. In this model, each observed variable is generated by a nonlinear function of its parents, with additive noise, followed by a nonlinear distortion. The nonlinearity in the second stage takes into account the effect of sensor distortions, which are usually encountered in practice. In the two-variable case, if all the nonlinearities involved in the model are invertible, by relating the proposed model to the post-nonlinear independent component analysis (ICA) problem, we give the conditions under which the causal relation can be uniquely found. We present a two-step method, which is constrained nonlinear ICA followed by statistical independence tests, to distinguish the cause from the effect in the two-variable case. We apply this method to solve the problem "CauseEffectPairs" in the Pot-luck challenge, and successfully identify causes from effects.

Keywords: causal discovery, sensor distortion, additive noise, nonlinear independent component analysis, independence tests

1. Introduction

Given some observable variables, people often wish to know the underlying mechanism generating them, and in particular, how they are influenced by others. Causal discovery has attracted much interest in various areas, such as philosophy, psychology, machine learning, etc. There are some well-known algorithms for causal discovery. For example, conditional independence tests can be exploited to remove unnecessary connections among the observed variables and to produce a set of acyclic causal models which are in the d-separation equivalence class (Pearl, 2000; Spirtes et al., 2000).

Recently, some methods have been proposed for model-based causal discovery of continuous variables (see, e.g., Shimizu et al., 2006; Granger, 1980). Model-based causal discovery assumes a generative model to explain the data generating process. If the assumed model is close to the true one, such methods could not only detect the causal relations, but also dis-

cover the form in which each variable is influenced by others. For example, Granger causality assumes that effects must follow causes and that the causal effects are linear (Granger, 1980). If the data are generated by a linear acyclic causal model and at most one of the disturbances is Gaussian, independent component analysis (ICA) (Hyvärinen et al., 2001) can be exploited to discover the causal relations in a convenient way (Shimizu et al., 2006).

However, the above causal models seem too restrictive for real-life problems. If the assumed model is wrong, model-based causal discovery may give misleading results. Therefore, when the prior knowledge about the data model is not available, the assumed model should be general enough such that it could be adapted to approximate the true data generating process. On the other hand, the model should be identifiable such that it could distinguish causes from effects. In a large class of real-life problems, the following three effects usually exist. 1. The effect of the causes is usually nonlinear. 2. The final effect received by the target variable from all its causes contains some noise which is independent from the causes. 3. Sensors or measurements may introduce nonlinear distortions into the observed values of the variables. To address these issues, we propose a very realistic model, called post-nonlinear acyclic causal model with inner additive noise. In the two-variable case, we show the identifiability of this model under the assumption that the involved nonlinearities are invertible. We conjecture that this model is identifiable in very general situations, as illustrated by the experimental results.

2. Proposed Causal Model

Let us use a directed acyclic graph (DAG) to describe the generating process of the observed variables. We assume that each observed continuous variable x_i, corresponding to the ith node in the DAG, is generated by two stages. The first stage is a nonlinear transformation of its parents pa_i, denoted by $f_{i,1}(pa_i)$, plus some noise (or disturbance) e_i (which is independent from pa_i). In the second stage, a nonlinear distortion $f_{i,2}$ is applied to the output of the first stage to produce x_i. Mathematically, the generating process of x_i is

$$x_i = f_{i,2}(f_{i,1}(pa_i) + e_i). \tag{1}$$

In this model, we assume that the nonlinearities $f_{i,2}$ are continuous and invertible. $f_{i,1}$ are not necessarily invertible. This model is very general, since it accounts for the nonlinear effect of the causes pa_i (by using $f_{i,1}$), the noise effect in the transmission process from pa_i to x_i (using e_i), and the nonlinear distortion caused by the sensor or measurement (using $f_{i,2}$). In particular, in this paper we focus on the two-variable case. Suppose that x_2 is caused by x_1. The relationship between x_1 and x_2 is then assumed to be

$$x_2 = f_{2,2}(f_{2,1}(x_1) + e_2), \tag{2}$$

where e_2 is independent from x_1.

3. Identifiability

3.1 Relation to post-nonlinear mixing ICA

We first consider the case where the nonlinear function $f_{2,1}$ is also invertible. Let $s_1 \triangleq f_{2,1}(x_1)$ and $s_2 \triangleq e_2$. As e_2 is independent from x_1, obviously s_1 is independent from s_2.

The generating process of (x_1, x_2), given by Eq. 2, can be re-written as

$$\begin{cases} x_1 = f_{2,1}^{-1}(s_1), \\ x_2 = f_{2,2}(s_1 + s_2). \end{cases} \tag{3}$$

We can see that clearly x_1 and x_2 are post-nonlinear (PNL) mixtures of independent sources s_1 and s_2 (Taleb and Jutten, 1999). The PNL mixing model is a nice special case of the general nonlinear ICA model.

ICA is a statistical technique aiming to recover independent sources from their observed mixtures, without knowing the mixing procedure or any specific knowledge of the sources (Hyvärinen et al., 2001). The basic ICA model is linear ICA, in which the observed mixtures, as components of the vector $\mathbf{x} = (x_1, x_2 \cdots, x_n)^T$, are assumed to be generated from the independent sources $s_1, s_2 \cdots, s_n$, with a linear transformation $\mathbf{A}$. Mathematically, we have $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s} = (s_1, s_2 \cdots, s_n)^T$. Under weak conditions on the source distribution and the mixing matrix, ICA can recover the original independent sources up to the permutation and scaling indeterminacies with another transformation $\mathbf{W}$, by making the outputs as independent as possible. That is, the outputs of ICA, as components of $\mathbf{y} = \mathbf{W}\mathbf{x}$, produce an estimate of the original sources s_i. In the general nonlinear ICA problem, $\mathbf{x}$ is assumed to be generated from independent sources s_i with an invertible nonlinear mapping $\mathcal{F}$, i.e., $\mathbf{x} = \mathcal{F}(\mathbf{s})$, and the separation system is $\mathbf{y} = \mathcal{G}(\mathbf{x})$, where $\mathcal{G}$ is another invertible nonlinear mapping. Generally speaking, nonlinear ICA is ill-posed: its solutions always exist but they are highly non-unique (Hyvärinen and Pajunen, 1999). To make the solution to nonlinear ICA meaningful, one usually needs to constrain the mixing mapping to have some specific forms (Jutten and Taleb, 2000).

The PNL mixing ICA model plays a nice trade-off of linear ICA and general nonlinear ICA. It is described as a linear transformation of the independent sources $s_1, s_2, ..., s_n$ with the transformation matrix $\mathbf{A}$, followed by a component-wise invertible nonlinear transformation $\mathbf{f} = (f_1, f_2, ..., f_n)^T$. Mathematically,

$$x_i = f_i\Big(\sum_{k=1}^{n} \mathbf{A}_{ik} s_k\Big).$$

In matrix form, it is denoted as $\mathbf{x} = \mathbf{f}(\mathbf{A}\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ and $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$. In particular, from Eq. 3, one can see that for the causal model Eq. 2, the mixing matrix is $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$, and the post-nonlinearity is $\mathbf{f} = (f_{2,1}^{-1}, f_{2,2})^T$.

3.2 Identifiability of the Causal Model

The identifiability of the causal model Eq. 2 is then related to the separability of the PNL mixing ICA model. The PNL mixing model $(\mathbf{A}, \mathbf{f})$ is said to be separable if the independent sources s_i could be recovered only up to some trivial indeterminacies (which includes the permutation, scaling, and mean indeterminacies) with a separation system $(\mathbf{g}, \mathbf{W})$, The output of the separation system is $\mathbf{y} = \mathbf{W} \cdot \mathbf{g}(\mathbf{x})$, where $\mathbf{g}$ is a component-wise continuous and invertible nonlinear transformation. The separability of the PNL mixing model has been discussed in several contributions. As Achard and Jutten (2005) proved the separability under very general conditions, their result is briefly reviewed below.

Theorem 1 (Separability of the PNL mixing model, by Achard & Jutten) *Let* $(\mathbf{A}, \mathbf{f})$ *be a PNL mixing system and* $(\mathbf{g}, \mathbf{W})$ *the separation system. Let* $h_i \triangleq g_i \circ f_i$. *Assume the following conditions hold.*

- *Each source* s_i *appears mixed at least once in the observations.*
- $h_1, h_2, ..., h_n$ *are diffenrentiable and invertible (same conditions as* $f_1, f_2, ..., f_n$*).*
- *There exists at most one Gaussian source.*
- *The joint density function of the sources* s_i *is differentiable, and its derivative is continuous on its support.*

Then the output of the separation system ($\mathbf{g}$,$\mathbf{W}$) *has mutually independent components if and only if each* h_i *is linear and* $\mathbf{WA}$ *is a generalized permutation matrix.*

The above theorem states that under the conditions stated above, by making the outputs of the separation system ($\mathbf{g}$, $\mathbf{W}$) mutually independent, the original sources s_i and the mixing matrix $\mathbf{A}$ could be uniquely estimated (up to some trivial indeterminacies). If $f_{2,1}$ is invertible, the causal model Eq. 2, as a special case of the PNL mixing model, can then be identified. Thus, the theorem above implies the following proposition.

Proposition 1 (Identifiability of the causal model with invertible nonlinearities) *Suppose that* x_1 *and* x_2 *are generated according to the causal model Eq. 2 with both* $f_{2,2}$ *and* $f_{2,1}$ *differentiable and invertible. Further assume that at most one of* $f_{2,1}(x_1)$ *and* e_2 *is Gaussian, and that their joint density is differentiable, with the derivative continuous on its support. Then the causal relation between* x_1 *and* x_2 *can be uniquely identified.*

In the discussions above, we have constrained the nonlinearity $f_{2,1}$ to be invertible. Otherwise, $f_{2,1}^{-1}$ does not exist, and the causal model Eq. 2 is no longer a PNL mixing one. A rigorous proof of the identifiability of the causal model in this situation is under investigation. But it seems that it is identifiable under very general conditions, as verified by various experiments. It should be noted that when all the nonlinear functions $f_{i,2}$ are constrained to be identity mappings, the proposed causal model is reduced to the nonlinear causal model with additive noise which was recently investigated by Hoyer et al. (2009). Interestingly, for this model, it was shown that in the two-variable case, the identifiability actually does not depend on the invertibility of the nonlinear function $f_{2,1}$.

4. Method for Identification

Given two variables x_1 and x_2, we identify their causal relation by finding which one of the possible relations ($x_1 \to x_2$ and $x_2 \to x_1$) satisfies the assumed causal model. If the causal relation is $x_1 \to x_2$ (i.e., x_1 and x_2 satisfy the model Eq. 2), we can invert the data generating process Eq. 2 to recover the disturbance e_2, which is expected to be independent from x_1. One can then examine if a possible causal model is preferred in two steps: the first step is actually a constrained nonlinear ICA problem which aims to retrieve the disturbance corresponding to the assume causal relation; in the second step we verify if the estimated disturbance is independent from the assume cause using statistical tests.

4.1 A two-step method

Suppose the causal relation under examination is $x_1 \to x_2$. According to Eq. 2, if this causal relation holds, there exist nonlinear functions $f_{2,2}^{-1}$ and $f_{2,1}$ such that $e_2 = f_{2,2}^{-1}(x_2) -$

$f_{2,1}(x_1)$ is independent from x_1. Thus, we first perform nonlinear ICA using the structure in Figure 1. The outputs of this system are $y_1 = x_1$, and $y_2 = g_2(x_2) - g_1(x_1)$. In our experiments, we use multi-layer perceptrons (MLP's) to model the nonlinearities g_1 and g_2. Parameters in g_1 and g_2 are learned by making y_1 and y_2 as independent as possible, which is achieved by minimizing the mutual information between y_1 and y_2. The joint density of $\mathbf{y} = (y_1, y_2)^T$ is $p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})/|\mathbf{J}|$, where $\mathbf{J}$ is the Jacobian matrix of the transformation from (x_1, x_2) to (y_1, y_2), i.e., $\mathbf{J} = \left[\partial(y_1, y_2)/\partial(x_1, x_2)\right]$. Clearly $|\mathbf{J}| = |g_2'|$. The joint entropy of $\mathbf{y}$ is then

$$H(\mathbf{y}) = -E\{\log p_{\mathbf{y}}(\mathbf{y})\} = -E\{\log p_{\mathbf{x}}(\mathbf{x}) - \log|\mathbf{J}|\} = H(\mathbf{x}) + E\{\log|\mathbf{J}|\}.$$

Finally, the mutual information between y_1 and y_2 is

$$\begin{aligned} I(y_1, y_2) &= H(y_1) + H(y_2) - H(\mathbf{y}) \\ &= H(y_1) + H(y_2) - E\{\log|\mathbf{J}|\} - H(\mathbf{x}) \\ &= -E\{p_{y_1}(y_1)\} - E\{p_{y_2}(y_2)\} - E\{\log|g_2'|\} - H(\mathbf{x}), \end{aligned}$$

where $H(\mathbf{x})$ does not depend on the parameters in g_1 and g_2 and can be considered as constant. One can easily find the gradient of $I(y_1, y_2)$ w.r.t. the parameters in g_1 and g_2, and minimize $I(y_1, y_2)$ using gradient-descent methods. Details of the algorithm are skipped.

y_1 and y_2 produced by the first step are the assumed cause and the estimated corresponding disturbance, respectively. In the second step, one needs to verify if they are independent, using statistical independence tests. We adopt the kernel-based statistical test (Gretton et al., 2008), with the significance level $\alpha = 0.01$. If y_1 and y_2 are not independent, indicating that $x_1 \to x_2$ does not hold, we repeat the above procedure (with x_1 and x_2 exchanged) to verify if $x_2 \to x_1$ holds. If y_1 and y_2 are independent, usually we can conclude that x_1 causes x_2, and that g_1 and g_2 provide an estimate of $f_{2,1}$ and $f_{2,2}^{-1}$, respectively. However, it is possible that both $x_1 \to x_2$ and $x_2 \to x_1$ hold, although the chance is very small. Therefore, for the sake of reliability, in this situation we also test if $x_2 \to x_1$ holds. Finally, we can find the relationship between x_1 and x_2 among all four possible scenarios: 1. $x_1 \to x_2$, 2. $x_2 \to x_1$, 3. both causal relations are possible, and 4. there is no causal relation between x_1 and x_2 which follows our model.

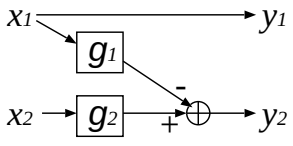

Figure 1: The constrained nonlinear ICA system used to verify if the causal relation $x_1 \to x_2$ holds.

4.2 Practical considerations

The first issue that needs considering in practical implementation of our method is the model complexity, which is controlled by the number of hidden units in the MLP's modelling g_1 and g_2 in Figure 1. The system should have enough flexibility, and at the same time, to avoid overfitting, it should be as simple as possible. To this end, two ways are used. One is 10-fold cross-validation. The other is heuristic: we try different numbers of hidden units in a reasonable range (say, between 4 and 10); if the resulting causal relation does not change, we conclude that the result is feasible.

Data Set	#1	#2	#3	#4	#5	#6	#7	#8
Result	$x_1 \to x_2$	$x_1 \to x_2$	$x_1 \to x_2$	$x_1 \overset{\ddagger}{\leftarrow} x_2$	$x_1 \leftarrow x_2$	$x_1 \to x_2$	$x_1 \leftarrow x_2$	$x_1 \to x_2$

Table 1: Causal directions obtained. (‡ indicates that the causal relation is not significant.)

The second issue is the initialization of the nonlinearities g_1 and g_2 in Figure 1. If the nonlinear distortions $f_{2,2}$ and $f_{2,1}$ are very strong, it may take a long time for the nonlinear ICA algorithm in the first step to converge, and it is also possible that the algorithm converges to a local optimum. This can be avoided by using reasonable initializations for g_1 and g_2. Two schemes are used in our experiments. One is motivated by visual inspection of the data distribution: we simply use a logarithm-like function to initialize g_1 and g_2 to make the transformed data more regular. The other is by making use of Gaussianization (Zhang and Chan, 2005). Roughly speaking, the central limit theorem states that sums of independent variables tend to be Gaussian. Since $f_{2,2}^{-1}(x_2)$ in the causal model Eq. 2 is the sum of two independent variables, it is expected to be not very far from Gaussian. Therefore, for each variable which is very far from Gaussian, its associated nonlinearity (g_1 or g_2 in Figure 1) is initialized by the strictly increasing function transforming this variable to standard Gaussian. In all experiments, these two schemes give the same final results.

5. Results

The proposed nonlinear causal discovery method has been applied to the "CauseEffectPairs" task proposed by Mooij and Janzing (2008) in the Pot-luck challenge. In this task, eight data sets are given; each of them contains the observed values of two variables x_1 and x_2. The goal is to distinguish the cause from the effect for each data set. Figure 2 gives the scatterplots of x_1 and x_2 in all the eight data sets. Table 1 summaries our results. In particular, below we take data sets 1 and 8 as examples to illustrate the performance of our method.

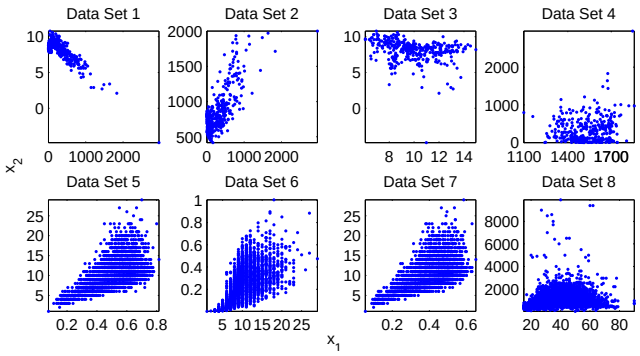

Figure 2: Scatterplot of x_1 and x_2 in each data set of the "CauseEffectPairs" task (Mooij and Janzing, 2008).

The variable x_1 in Data set 1 is non-negative and extremely non-Gaussian. We initialized the nonlinearity g_1 with the transformation $\log(2+x_1)$ (Gaussianization was also tested and it finally produced the same causal relation). The scatterplot of y_1 and y_2 (as outputs of the constrained nonlinear ICA system in Figure 1) under each hypothesis ($x_1 \to x_2$ or $x_2 \to x_1$) is given in Figure 3(a,b). Clearly y_1 and y_2 are much more independent under hypothesis $x_1 \to x_2$. This is verified by the independence test results in the third row of Table 2. Note that a large test statistic tends to reject the null hypothesis (the independence between y_1 and y_2). Figure 4 shows the result on Data set 8. In this case, we applied the transformation $\log(x_2 + 50)$ for initialization. By comparing (a) and (b) in Figure 4, also by inspecting the independence test results in the fourth row of Table 2, one can see clearly that $x_1 \to x_2$.

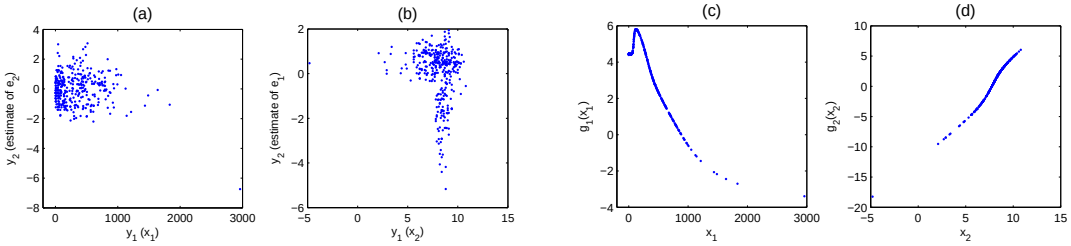

Figure 3: Result on Data set 1. (a) y_1 vs. y_2 under hypothesis $x_1 \to x_2$. (b) that under $x_2 \to x_1$. (c & d) x_1 vs. $g_1(x_1)$ and x_2 vs. $g_2(x_2)$ under hypothesis $x_1 \to x_2$.

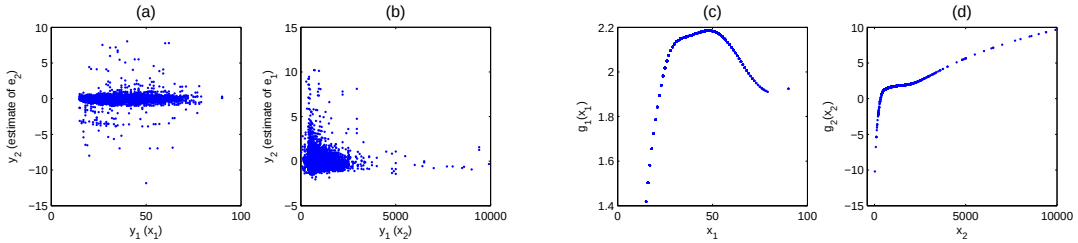

Figure 4: Result on Data set 8. For captions of the sub-figures, please refer to Figure 3.

Data Set	$x_1 \to x_2$ assumed		$x_2 \to x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}
#8	1.2×10^{-4}	1.2×10^{-4}	1.1×10^{-4}	7.4×10^{-4}

Table 2: Result of independence test on y_1 and y_2 for Data sets 1 and 8 under different assumed causal directions. For both data sets, the independence hypothesis is accepted in the scenario $x_1 \to x_2$, and rejected in the other scenario, with the significance level $\alpha = 0.01$.

6. Conclusion

We proposed a very general nonlinear causal model for model-based causal discovery. This model takes into account the nonlinear effect of the causes, inner noise effect, and the sensor distortion, and is capable of approximating the data generating process of some real-life problems. We presented the identifiability of this model under the assumption that the involved nonlinearities are invertible. Experimental results illustrated that based on this model, one could successfully distinguish the cause from the effect, even if the nonlinear function of the cause is not invertible. An on-going work is to investigate the identifiability of this model under more general conditions.

References

S. Achard and C. Jutten. Identifiability of post-nonlinear mixtures. *IEEE Signal Processing Letters*, 12:423–426, 2005.

C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 1980.

A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A.J. Smola. A kernel statistical test of independence. In *NIPS 20*, pages 585–592, Cambridge, MA, 2008.

P.O. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS 21*, Vancouver, B.C., Canada, 2009. To appear.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

C. Jutten and A. Taleb. Source separation: From dusk till dawn. In *2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, pages 15–26, Helsinki, Finland, 2000.

J. Mooij and D. Janzing. Distinguishing between cause and effect, Oct. 2008. URL `http://www.kyb.tuebingen.mpg.de/bs/people/jorism/causality-data/`.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition edition, 2000.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

K. Zhang and L. W. Chan. Extended Gaussianization method for blind separation of post-nonlinear mixtures. *Neural Computation*, 17(2):425–452, 2005.