Causality Discovery with Additive Disturbances: An Information-Theoretical Perspective

Kun Zhang[1] and Aapo Hyvärinen[1,2]

[1] Dept of Computer Science & HIIT, University of Helsinki, Finland
[2] Dept of Mathematics and Statistics, University of Helsinki, Finland

Abstract. We consider causally sufficient acyclic causal models in which the relationship among the variables is nonlinear while disturbances have linear effects, and show that three principles, namely, the causal Markov condition (together with the independence between each disturbance and the corresponding parents), minimum disturbance entropy, and mutual independence of the disturbances, are equivalent. This motivates new and more efficient methods for some causal discovery problems. In particular, we propose to use multichannel blind deconvolution, an extension of independent component analysis, to do Granger causality analysis with instantaneous effects. This approach gives more accurate estimates of the parameters and can easily incorporate sparsity constraints. For additive disturbance-based nonlinear causal discovery, we first make use of the conditional independence relationships to obtain the equivalence class; undetermined causal directions are then found by nonlinear regression and pairwise independence tests. This avoids the brute-force search and greatly reduces the computational load.

1 Introduction

Given some observed variables, scientists, engineers, and policy-makers often wish to find their causal relations, as well as to understand how to control a particular variable by manipulating others. Discovering causal relations from non-experimental data has attracted the interests of researchers in many areas, such as philosophy, psychology, machine learning, etc [13, 19]. In this paper we focus on the causally sufficient acyclic causal models [13] of continuous variables. That is, we assume that there are no confounders nor any feedback in the causal relations.

There are some frequently-used models for acyclic causal discovery, and traditionally they are estimated with different principles. For example, for Gaussian variables with linear relations, conditional independence between the variables allows one to find a set of acyclic causal models which are in the d-separation equivalence class [13]. Generally speaking, with more specific information about the disturbance distribution or the causal structure, one can find the underlying causal model more accurately. Based on the independent component analysis (ICA [8]) technique, a class of linear, non-Gaussian, and acyclic models (LiNGAM) can be estimated very efficiently [18]. Moreover, in economics, Granger causality analysis [4] is a popular way to examine the causal relations between times series. It exploits the temporal constraint that causes must precede effects and uses the vector auto-regression (VAR) for parameter estimation.

In this paper, we consider a large class of acyclic causal models in which the causal relations among observed variables are nonlinear but the effect of disturbances is linear, as extensions of the linear models mentioned above. We show that for such causal models, mutual independence of the disturbances is equivalent to conditional independence of observed variables (as well as the independence between the disturbance and the parents affecting the same variable). Furthermore, they are achieved if and only if the total entropy of the disturbances is minimized. The three criteria above, namely, conditional independence of variables (together with the independence between each disturbance and the corresponding parents), mutual independence of disturbances, and minimum disturbance entropy, can all be exploited to estimate such causal models. In practice, which one should be chosen depends on the problem at hand.

We then consider two causal discovery problems, and show how our results help solve them efficiently. One is discovery of Granger causality with instantaneous effects. Previous methods consist of two separate steps: the first step performs ordinary Granger causality analysis by using VAR's, and the second step finds the instantaneous causal relations [16, 10]. Although these methods are consistent in large samples, they are not efficient, because the Gaussianity assumption for the innovations made in the first step is usually not true. We propose a more efficient approach to estimate this model by making the disturbances mutually independent, as achieved by multichannel blind deconvolution (MBD) [1], an extension of ICA.

The second problem on which we apply our theory is nonlinear causal discovery with additive disturbances in the case of more than two variables. The existing approach requires an exhaustive search over all possible causal structures and testing if the disturbances are mutually independent [6]. It becomes impractical when we have more than three or four variables. The proposed approach, which can easily solve this problem with tens of variables, consists of two stages. First, using nonlinear regression and statistical independence tests, one can find the *d*-separation equivalence class. Next, among all possible causal models in the equivalence class, the one consistent with the data can be found by examining if the disturbance is independent of the parents for each variable.

2 Equivalence of Three Estimation Principles for Acyclic Causal Models

In this section we consider a kind of acyclic data generating processes in which each variable is generated by a nonlinear function of its parents plus the disturbance. Such processes can be represented graphically by a directed acyclic graph (DAG). Mathematically, each of the observed variables x_i, $i = 1, \cdots, n$, is written as

$$x_i = f_i(pa_i) + e_i, \tag{1}$$

where pa_i denotes the parents of x_i, and the disturbances e_i are independent from each other.

A well-known approach to identify the acyclic causal relations is based on a test called d-separation, which examines conditional independence between variables [13, 19]. In the following theorem, we show that mutual independence of the disturbances and conditional independence between observed variables (together with the independence between e_i and pa_i) are equivalent. Furthermore, they are achieved if and only if the total entropy of the disturbances is minimized.

Theorem 1 *Assume that the data $x_1, \cdots, x_n$ are causally sufficient, and were generated by the acyclic causal model in Eq. 1. Then, when fitting the model Eq. 1 with the causal structure represented by a DAG to $x_1, \cdots, x_n$, the following three properties are equivalent:*

(i) *The causal Markov condition holds (i.e., each variable is independent of its non-descendants in the DAG conditional on its parents), and in addition, the disturbance in x_i is independent from the parents of x_i.*[3]
(ii) *The total entropy of the disturbances, i.e., $\sum_i H(e_i)$, is minimized, with the minimum $H(x_1, \cdots, x_n)$.*
(iii) *The disturbances e_i are mutually independent.*

See Appendix for a proof. From this theorem, one can see that $\sum_{i=1}^n H(e_i)$ can be used as a measure to compare the quality of different acyclic causal model. Generally, we prefer the model with the smallest $\sum_{i=1}^n H(e_i)$. This is intuitively appealing from a physics viewpoint. In physics, it is often claimed that causality increases entropy [3]. Therefore, the disturbances e_i, which are pure causes of the observed variables x_i, should have the least total entropy, compared to any observed variables they generate. Another interesting point is that although the causal relations among x_i increases the total entropy of the observed variables, i.e., $\sum_i H(x_i)$, the joint entropy of the x_i remains the same, with the value $\sum_i H(e_i)$. In addition, the property (ii) for the acyclic causal model relates causality and predictability from an information theory viewpoint. The causal relations give the best prediction all variables in the system in an acyclic manner, since the uncertainty (measured by entropy) in all errors e_i is minimized.

All the three criteria in Theorem 1 can be used for causal discovery. However, for a given problem, they may result in different techniques and involve different computational loads. It should be noted that when using the minimum disturbance entropy criterion, one needs to constrain the causal models to be acyclic; otherwise, minimization of this criterion tends to introduce cyclic relations into the model, to reduce the magnitude of the disturbances.

It should be noted that the theorem does not state when the solution of the problem is unique. It is well-known that for Gaussian variables, there may be several solutions providing independent disturbances, and thus minimum entropy.

[3] The property that the disturbance in x_i is independent from the parents of x_i is trivial in the linear Gaussian case, since using linear regression, the noise (disturbance) is uncorrelated from the explanatory variables (parents), and uncorrelatedness is equivalent to independence for jointly Gaussian variables.

Finding exact conditions for uniqueness is an important problem for future research. For example, non-Gaussianity, or temporal order, have been shown to make the solution unique in the linear case [18, 4].

3 On Granger Causality with Instantaneous Effects

Granger causality [4] exploits the temporal information that the cause occurs before the effect, such that the directions of the possible causal effects are known. Consequently, it can be determined uniquely. A stationary process $X_1 : \{x_{1t}\}$ is said to Granger cause the process $X_2 : \{x_{2t}\}$ if it contains information about the predictability for $x_{2,t+1}$ contained nowhere else in some large information set, which includes $x_{1,t-k}, k \geq 0$ [4]. Using the language of conditional independence, this means that X_1 is not conditionally independent of $x_{2,t+1}$ given the large information set. In this sense, Granger causality is a kind of conditional independence-based causality combined with the constraint that effects must follow causes. Here we would like to treat $x_{i,t}$ as random variables, and use a DAG to represent the possible causal relations among them. If there exist significant causal relations $x_{i,t-k} \rightarrow x_{j,t}$ ($k > 0$, and $i \neq j$), then the process X_i Granger causes X_j.

As the causal relations among $x_{i,t}$ are linear and acyclic, the three properties in Theorem 1 can all be used to estimate this model. Conventionally, all of $e_{i,t}$ are treated as Gaussian, and minimization of the disturbance entropy is reduced to minimizing the prediction error (in the mean square error sense), as achieved by VAR's. It should be noted that when e_{it} are not Gaussian, although the estimate given by VAR's is consistent in large samples, it is not efficient. With Granger causality analysis, it is sometimes observed that there is significant contemporaneous dependency between the innovations e_i. This means that there are some instantaneous relations among x_{it}, which cannot be captured by traditional Granger causality.

3.1 Granger Causality with Instantaneous Effects

Mathematically, Granger causality analysis of $x_{1t}, \cdots, x_{nt}$ with instantaneous effects can be formulated as

$$\mathbf{x}_t = \sum_{\tau=0}^{p} \mathbf{B}_\tau \mathbf{x}_{t-\tau} + \mathbf{e}_t, \tag{2}$$

where $\mathbf{x}_t = (x_{1t}, \cdots, x_{nt})^T$, $\mathbf{e}_t = (e_{1t}, \cdots, e_{nt})^T$, and $\mathbf{B}_\tau$ are $n \times n$ matrix of coefficients. Here we have assumed that all involved random variables have been made zero-mean. We also assume that the instantaneous causal relations, which are implied in $\mathbf{B}_0$, are acyclic. That is, $\mathbf{B}_0$ can be transformed to a strictly lower-triangular matrix by simultaneous equal row and column permutations [18]. Equivalently, the representation Eq. 2 can be written as

$$(\mathbf{I} - \mathbf{B}_0)\mathbf{x}_t = \sum_{\tau=1}^{p} \mathbf{B}_\tau \mathbf{x}_{t-\tau} + \mathbf{e}_t. \tag{3}$$

3.2 Existing Methods

Existing methods for Granger causality analysis with instantaneous effects consist of two steps [10, 16]. Multiplying both sides of Eq. 3 by $(\mathbf{I}-\mathbf{B}_0)^{-1}$ from the left, one can get

$$\mathbf{x}_t = \sum_{\tau=1}^{p} (\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{B}_\tau \cdot \mathbf{x}_{t-\tau} + (\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{e}_t. \tag{4}$$

This is exactly the Granger causality model without instantaneous effects with the errors $(\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{e}_t$. Therefore, one can first find $(\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{B}_\tau$, $\tau = 1, \cdots, p$, and $(\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{e}_t$, by ordinary VAR analysis. In the second step, one needs to estimate $\mathbf{B}_0$ by examining (the estimate of) the errors $(\mathbf{I}-\mathbf{B}_0)^{-1} \cdot \mathbf{e}_t$. One way is based on conditional independence graphs [16]. It is a combination of VAR for lagged causality and conditional independence-based method [13] for instantaneous causality. Due to the Gaussianity assumption, this method produces a distribution-equivalence class of causal models. Another way, recently proposed in [10], resorts to the ICA-based LiNGAM analysis [18]. The method is very easy to implement, and also consistent in large samples. But when at most one of the disturbance sequence is Gaussian, it is not efficient due to the wrong assumption of Gaussianity of the disturbances in the first step and the error accumulation of the two-step method.

3.3 Estimation by Multichannel Blind Deconvolution

According to the causal model Eq. 2, the causal relations among random variables x_{it} are linear and acyclic. According to Theorem 1, such a causal model can be estimated by making the disturbances e_{it} mutually independent for different i and different t. That is, we need to make e_{it}, which are a mixed and filtered version of $\mathbf{x}_t$, both spatially and temporally independent. Estimation of the model Eq. 3 (or equivalently, Eq. 2) is then closely related to the multichannel blind deconvolution (MBD) problem with causal finite impulse response (FIR) filters [1, 8]. MBD, as a direct extension of ICA [8], assumes that the observed signals are convolutive mixtures of some spatially and independently and identically distributed (i.i.d.) sources. Under the assumption that at most one of the sources is Gaussian, by making the estimated sources spatially and temporally independent, MBD can recover the mixing system (here corresponding to e_{it} and $\mathbf{B}_\tau$) up to some scaling, permutation, and time shift indeterminacies [11]. This implies that Granger causality with instantaneous effects is identifiable if at most one of the disturbances e_i is Gaussian.

In Eq. 2, the observed variables x_{it} can be considered as convolutive mixtures of the disturbances e_{it}. We aim to find the estimate of $\mathbf{B}_\tau$, as well as e_{it}, in Eq. 2, by MBD with the filter matrix $\mathbf{W}(z) = \sum_{\tau=0}^{p} \mathbf{W}_\tau z^{-\tau}$ ($\mathbf{W}_\tau$ are $n \times n$ matrices):

$$\hat{\mathbf{e}}_t = \sum_{\tau=0}^{p} \mathbf{W}_\tau \mathbf{x}_{t-\tau}. \tag{5}$$

There exist several well-developed algorithms for MBD. For example, one may adopt the one based on natural gradient [1]. Comparing Eq. 5 and Eq. 3, one can see that the estimate of $\mathbf{B}_\tau$ ($\tau \geq 0$) can be constructed by analyzing $\mathbf{W}_\tau$: by extending the LiNGAM analysis procedure [18], we can find the estimate of $\mathbf{B}_\tau$ in the following three steps, based on the MBD estimates of $\mathbf{W}_\tau$.

1. Find the permutation of rows of $\mathbf{W}_0$ which yields a matrix $\widetilde{\mathbf{W}}_0$ without any insignificant entries on the main diagonal. Note that here we also need to apply the same permutations to rows of $\mathbf{W}_\tau$ ($\tau > 0$) to produce $\widetilde{\mathbf{W}}_\tau$.
2. Divide each row of $\widetilde{\mathbf{W}}_0$ and $\widetilde{\mathbf{W}}_\tau$ ($\tau > 0$) by the corresponding diagonal entry in $\widetilde{\mathbf{W}}_0$. This gives $\widetilde{\mathbf{W}}'_0$ and $\widetilde{\mathbf{W}}'_\tau$, respectively. The estimate of $\mathbf{B}_0$ and $\mathbf{B}_\tau$ ($\tau > 0$) can be computed as $\widehat{\mathbf{B}}_0 = \mathbf{I} - \widetilde{\mathbf{W}}'_0$ and $\widehat{\mathbf{B}}_\tau = -\widetilde{\mathbf{W}}'_\tau$, respectively.
3. To obtain the causal order in the instantaneous effects, find the permutation matrix $\mathbf{P}$ (applied equally to both rows and columns) of $\widehat{\mathbf{B}}_0$ which makes $\widetilde{\mathbf{B}}_0 = \mathbf{P}\widehat{\mathbf{B}}_0\mathbf{P}^T$ as close as possible to strictly lower triangular.

3.4 Sparsification of the Causal Relations

For the interpretation or generalization purpose, we need to do model selection on the causal structure (i.e., to set insignificant entries of $\hat{\mathbf{B}}_\tau$ to zero, and to determine p, if needed). This is difficult to do in the two-step methods [16, 10], but is easily achieved in our method. Analogously to the development of ICA with sparse connections [20], we can incorporate the adaptive L_1 penalties into the likelihood of the MBD model to achieve fast model selection. More importantly, the model selection result obtained by this approach is consistent with that by traditional information criteria, such as BIC [17].[4] To make $\mathbf{W}_\tau$ in Eq. 5 as sparse as possible, we maximize the penalized likelihood

$$pl(\{\mathbf{W}_\tau\}) = l(\{\mathbf{W}_\tau\}) - \lambda \sum_{i,j,\tau} |w_{i,j,\tau}|/|\hat{w}_{i,j,\tau}|, \tag{6}$$

where $l(\{\mathbf{W}_\tau\})$ is the likelihood, $w_{i,j,\tau}$ the (i,j)th entry of $\mathbf{W}_\tau$, and $\hat{w}_{i,j,\tau}$ a consistent estimate of $w_{i,j,\tau}$, such as the maximum likelihood estimate. To achieve BIC-like model selection, one can set $\lambda = \log T$, where T is the sample size.

3.5 Simulation

To investigate the performance of our method, we conducted a series of simulations. We set $p = 1$ lag and the dimensionality $n = 5$. We randomly constructed the strictly lower-triangular matrix $\mathbf{B}_0$ and matrix $\mathbf{B}_1$. About 60% of the entries

[4] Note that traditional model selection based on the information criteria involves a combinatorial optimization problem, whose complexity increases exponentially in the dimensionality of the parameter space. In the MBD problem, it is not practical to set insignificant entries of $\hat{\mathbf{B}}_\tau$ to zero by directly minimizing the information criteria, as the number of parameters is too large.

in these matrices were set to zero, while the magnitude of the others is uniformly distributed between 0.05 and 0.5 and the sign is random. The disturbances e_{it} were generated by passing i.i.d. Gaussian samples through a power nonlinearity with exponent between 1.5 and 2.0 (the original sign was kept). The observations $\mathbf{x}_t$ were then generated according to Eq. 4. Various sample sizes ($T = 100$, 300, and 1000) were tested. We compared the performance of the two-step method proposed in [10], the method by MBD (Section 3.3) and the MBD-based method with the sparsity constraint (Section 3.4). In the last method, we set the penalization parameter in Eq. 6 as $\lambda = \log T$ to make its results consistent with those obtained by BIC. In each case, we repeated the experiments for 5 replications.

Fig. 1 shows the scatter plots of the estimated parameters (including the strictly lower triangular part of $\mathbf{B_0}$ and all entries of $\mathbf{B}_1$) versus the true ones. Different subplots correspond to different sample sizes or different methods. The mean square error (MSE) of the estimated parameters is also given in each subplot. One can see that as the sample sizes increases, all methods give better results. For each sample size, the method based on MBD is always better than the two-step method, showing that the estimate by the MBD-based method is more efficient. Furthermore, due to the prior knowledge that many parameters are zero, the MBD-based method with the sparsity constraint behaves best.

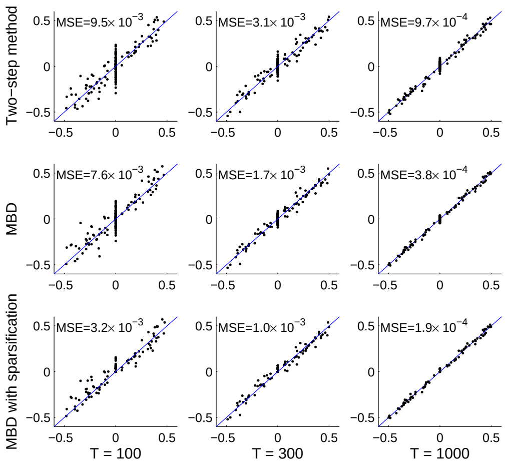


Fig. 1. Scatter plots of the estimated coefficients (y axis) versus the true ones (x axis) for different sample sizes and different methods.

3.6 Application in Finance

In this section we aim at using Granger causality analysis with instantaneous effects to find the causal relations among several world stock indices. The chosen indices are Dow Jones Industrial Average (DJI) in USA, Nikkei 225 (N225) in Japan, Hang Seng Index (HSI) in Hong Kong, and the Shanghai Stock Exchange Composite Index (SSEC) in China. We used the daily dividend/split adjusted closing prices from Dec. 4, 2001 to Jul. 11, 2006, obtained from the Yahoo finance database. For the few days when the price is not available, we use simple linear interpolation to estimate the price. Denoting the closing price of the ith index on day t by P_{it}, the corresponding return is calculated by $x_{it} = \frac{P_{it} - P_{i,t-1}}{P_{i,t-1}}$. The data for analysis are $\mathbf{x}_t = [x_{1t}, ..., x_{4,t}]^T$, with 1200 samples.

We applied the MBD-based method with the sparsity constraint to $\mathbf{x}_t$. The kurtoses of the estimated disturbances $\hat{e}_{it}$ are 3.9, 8.6, 4.1, and 7.6, respectively, implying that the disturbances are non-Gaussian. We found that more than half of the coefficients in the estimated $\mathbf{W}_0$ and $\mathbf{W}_1$ are zero. $\widehat{\mathbf{B}}_0$ and $\widehat{\mathbf{B}}_0$ were constructed based on $\mathbf{W}_0$ and $\mathbf{W}_1$, using the procedure given in Section 3.3. It was found that $\widehat{\mathbf{B}}_0$ can be permuted to a strictly lower-triangular matrix, meaning that the instantaneous effects follow a linear acyclic causal model. Finally, based on $\widehat{\mathbf{B}}_0$ and $\widehat{\mathbf{B}}_1$, one can plot the causal diagram, as shown in Fig. 2.

Fig. 2 reveals some interesting findings. First, DJI_{t-1} has significant impacts on N225_t and HSI_t, which is a well-known fact in the stock market. Second, the causal relations $\text{DJI}_{t-1} \rightarrow \text{N225}_t \rightarrow \text{DJI}_t$ and $\text{DJI}_{t-1} \rightarrow \text{HSI}_t \rightarrow \text{DJI}_t$ are consistent with the time difference between Asian and USA. That is, the causal effects from N225_t and HSI_t to DJI_t, although seeming to be instantaneous, are actually mainly caused by the time difference. Third, unlike SSEC, HSI is very sensitive to others; it is even strongly influenced by N225, another Asian index. Fourth, it may be surprising that there is a significant negative effect from DJI_{t-1} to DJI_t; however, it is not necessary for DJI_t to have significant negative autocorrelations, due to the positive effect from DJI_{t-1} to DJI_t going through N225_t and HSI_t.

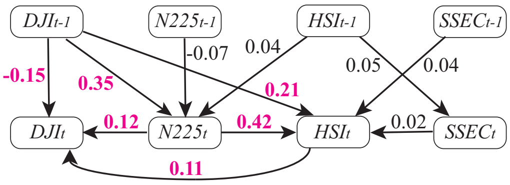


Fig. 2. Result of applying Granger causality analysis with instantaneous effects to daily returns of the stock indices DJI, N225, HSI, and SSEC, with $p = 1$ lag. Large coefficients (greater than 0.1) are shown in bold.

4 Additive Disturbance-Based Nonlinear Causal discovery with More than Two Variables

The additive disturbance causal model Eq. 1 has been proposed for nonlinear causal discovery very recently by Hoyer et al. [6]. They mainly focused on the two-variable case and showed that the model is able to distinguish the cause from effect for some real-world data sets. Suppose we have two variables x_1 and x_2. The method to find the causal relation between them is as follows. First, examine if they are independent using statistical independence tests, such as the kernel-based method [5]. If they are, no further analysis is needed. Otherwise, one then continues by testing if the model $x_2 = f_2(x_1) + e_2$ is consistent with the data. Nonlinear regression is used to find the estimate of f_2 and e_2. If $\hat{e}_2$ is independent of x_1, the causal model $x_2 = f_2(x_1) + e_2$ is accepted; otherwise it is rejected. One then needs to test if the reverse model $x_1 = f_1(x_2) + e_1$ is acceptable. Finally, if one model is accepted and the other is rejected, the additive disturbance nonlinear causal model is uniquely found. If both models hold, one can conclude that in this situation the additive disturbance causal model cannot distinguish the cause from effect. If neither of the models holds, one can conclude that the model cannot explain the data well, possibly due to the existence of hidden variables or a different data generating process. When performing nonlinear regression, one should carefully avoid over-fitting.

4.1 Existing Methods

It is much more difficult to find the causal relations implied by the causal model Eq. 1 when we have more than two observed variables. In this case, a brute-force search was exploited in [5]; for each possible acyclic causal structure, represented by a DAG, one performs nonlinear regression of each variable on its parents, and tests if the residuals are *mutually* independent with statistical independence tests. The simplest causal model which gives mutually independent disturbances is preferred. Clearly this approach may encounter two difficulties. One is that the test of mutual independence is difficult to do when we have many variables. The other is that the search space of all possible DAG's increases too rapidly with the variable number. In fact, it is well-known that the total number of all possible DAG's is super-exponential in the number of variables. Consequently, this approach involves high computational load, and is not suitable when we have more than three or four variables.

4.2 A More Practical Approach

We propose an approach which is suitable for identifying the nonlinear causal model Eq. 1 with moderate-sized variables, say, with tens of variables. According to Theorem 1, the nonlinear causal model Eq. 1 can be identified by enforcing the causal Markov property and the independence between the disturbance and the parents associated with the same variable. This motivates a two-stage approach to identify the whole causal model. One can first use conditional independence-based methods to find the d-separation equivalence class. Next, the nonlinear causal model Eq. 1 is used to identify the causal relations that cannot be determined in the first step: for each possible causal model contained in the equivalence class, we estimate the disturbances, and determine if this model is plausible, by examining if the disturbance in each variable x_i is independent of the parents of x_i.[5] In this way, one avoids the exhaustive search over all possible causal structures and statistical tests of mutual independence of more than two variables.

In the first stage of the proposed approach, we need to find an efficient way to derive the conditional independence relationships and to construct the equivalence class. This brings up two issues. One is how to reliably test conditional independence for variables with nonlinear relations. Traditionally, in the implementation of most conditional independence-based causal discovery algorithms, such as PC [19], it is assumed that the variables are either discrete or Gaussian with linear causal relations. This assumption greatly simplifies the difficulty in conditional independent tests. But here we need to capture the nonlinear causal effects. Some methods, such as the probabilistic non-parametric test proposed in [12], have been developed for this task. However, they may be unreliable when the conditional set contains many variables, due to the curse of dimensionality. Alternatively, one may simplify the conditional independence test procedure by making use of the particular structure of the nonlinear causal model Eq. 1. This can be done by extending the partial correlation concept to the nonlinear case.

Partial correlation measures the degree of linear association between two variables, with the effect of a set of conditional variables removed. In particular, the partial correlation between X and Y given a set of variables $\mathbf{Z}$, denoted by $\rho_{XY\cdot\mathbf{Z}}$, is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with $\mathbf{Z}$ and of Y with $\mathbf{Z}$, respectively. Here we assume that the data follow the nonlinear generating process Eq. 1. Due to the additive disturbance structure, one can examine if X and Y are conditionally independent given the variable set $\mathbf{Z}$ by performing independent tests on the residuals $R_X^{\mathcal{N}}$ and $R_Y^{\mathcal{N}}$, which are obtained by *nonlinear* regression of X with $\mathbf{Z}$ and of Y with $\mathbf{Z}$, respectively. In our implementation, Gaussian process regression with a Gaussian kernel [15] is adopted for nonlinear regression, and the involved hyperparameters are learned by maximizing the marginal likelihood. The kernel-based independence test [5] with the significance level 0.01 is then used to test if the residuals are independent.

The other issue is how to construct the independence-based equivalence class with as few conditional independence tests as possible. We adopt the total conditioning scheme discussed in [14], which was shown to be very efficient provided the underlying graph is sparse enough. It first finds the Markov blanket of each variable and builds the moral graph. The Markov blanket of the variable X is

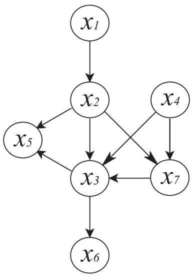
[5] According to Theorem 1, one can use the total entropy of the disturbances as the criterion to find the "best" causal model in the equivalence class. However, this approach does not easily provide a criterion for testing model validity.

the set of parents, children, and children's parents (spouses) of X. Let $\mathbf{V}$ be the set of all variables. The variable Y is in the Markov blanket of X if X and Y are not conditionally independent given $\mathbf{V} \setminus \{X, Y\}$. In particular, in our case we use nonlinear regression combined with the kernel-based independence test to perform the conditional independence test, as discussed above. Next, it removes the possible spouse links between linked variables X and Y by looking for a d-separating set around X and Y. When spouse links are removed, the V-structures can be oriented using collider sets. One can then find the equivalence class by propagating orientation constraints. For details of the total conditioning scheme for causal discovery based on Markov blankets, see [14]. In order to construct the moral graph, one needs to perform nonlinear regressions for $n(n-1)$ times and independence tests for $\frac{n(n-1)}{2}$ times, where n is the number of observed variables. To find the possible spouse links, one further needs to do nonlinear regressions for $2^{\alpha+1}$ times and corresponding independence tests for 2^{α} times, where $\alpha = \max_{X,Y} |\mathbf{Tri}(X - Y)|$, with $\mathbf{Tri}(X - Y)$ denoting the set of variables forming a triangle with X and Y.

Finally, for each DAG in the equivalence class, we use nonlinear regression to estimate the disturbances and then test if the disturbance and parents are independent for each variable. Those that make each disturbance independent of the parents associated with the same variable are valid models.

4.3 Simulation

In this section we investigate how the proposed approach behaves with a simulation study. The data generating process is given in Fig. 3. It consists of seven variables with both linear and strongly nonlinear causal relations, and the disturbances are Gaussian, uniform, or super-Gaussian. The sample size is 1000.



Data-generating process:
$x_1 = v_1$, with normal v_1;
$x_2 = x_1/2 + x_1^3/10 + v_2$, with normal v_2;
$x_4 = v_4$, with normal v_4;
$x_3 = x_2 + x_4 + v_3$, with super-Gaussian v_3;
$x_5 = x_2 x_3 + v_5$, with uniform v_5;
$x_6 = x_3/2 + x_3^3/10 + v_6$, with normal v_6;
$x_7 = \|x_3\|^{1.3}(1 + x_2 + \tanh(x_4)) + v_7$, with normal v_7.

Fig. 3. The true data-generating model used in the simulation, where v_i are mutually independent, the standard deviation of v_i is a random number between 0.2 and 1, and v_3 is obtained by passing a Gaussian variable through the power nonlinearity with exponent 1.5.

We first used nonlinear regression and independence test to construct the moral graph, with the result shown in Fig. 4(a). One can see that it is exactly the moral graph corresponding to the causal model generating the data. The edge $x_2 - x_4$ was then found to be a spouse link, since they are (unconditionally) independent. Consequently, x_3 and x_7 are colliders and thus common children of x_2 and x_4. That is, we have $x_2 \rightarrow x_3 \leftarrow x_4$ and $x_2 \rightarrow x_7 \leftarrow x_4$. Furthermore, since x_5 (x_6) is not connected to x_4 in the moral graph, one can find the orientation $x_3 \rightarrow x_5$ ($x_3 \rightarrow x_6$). To avoid cyclicity, the causal direction between x_2 and x_5 must be $x_2 \rightarrow x_5$. The resulting equivalence class is shown in Fig. 4(b), with only the causal direction between x_1 and x_2 and that between x_3 and x_7 are not determined. Under the hypothesis $x_3 \leftarrow x_7$, we found that the disturbance estimated by nonlinear regression is independent of the assumed parents (x_2, x_4, and x_7), while the disturbance is not independent of the parents for the variable x_7 under the hypothesis $x_3 \rightarrow x_7$, so we obtained $x_3 \leftarrow x_7$. Similarly, one can find the causal direction $x_1 \rightarrow x_2$. The obtained causal model is given in Fig. 4(c), which turns out to be the same as the one generating the data (Fig. 3). For comparison, we also show the equivalence class obtained by the PC algorithm [19] implemented in Tetrad [6] with the significance level 0.01. One can see that since the linearity assumption in PC is violated in this case, the resulting equivalence class is significantly different from the true causal model; in fact, half of the edges are spurious.

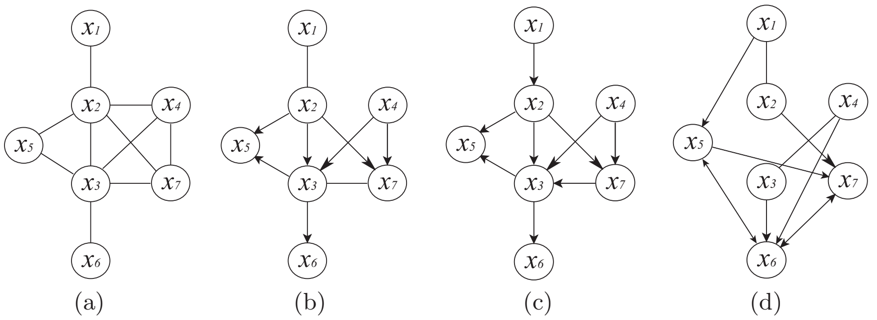


Fig. 4. Simulation results of nonlinear causal discovery with additive disturbances. (a) The moral graph obtained by our approach (the significance level for independence tests is 0.01). (b) The graph after removing spouse link ($x_2 - x_4$), orienting V-structures, and propagating orientation constraints. (c) The final result. (d) For comparison, the result obtained by the PC algorithm, with the significance level 0.01.

4.4 Application in Causal Discovery of MEG Data

As an illustration of the applicability of the method on real data, we applied it on magnetoencephalography (MEG), i.e., measurements of the electric activity in the brain. The raw data consisted of the 306 MEG channels measured by the Vectorview helmet-shaped neuromagnetometer (Neuromag Ltd., Helsinki, Finland)

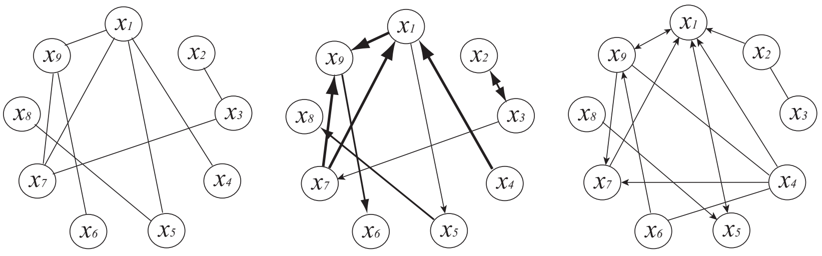
[6] Available at http : //www.phil.cmu.edu/projects/tetrad/

in a magnetically shielded room at the Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology. The measurements consisted of 300 seconds of resting state brain activity earlier used in [9]. The subject was sitting with eyes closed, and did not perform any specific task nor was there any specific sensory stimulation.

As pre-processing, we performed a blind separation of sources using the method called Fourier-ICA [9]. This gave nine sources of oscillatory brain activity. Our goal was to analyze the causal relations between the powers of the source, so we divided the data into windows of length of one second (half overlapping, i.e., the initial points were at a distance of 0.5 seconds each) and computed the logarithm of the local standard deviation in each window. This gave a total of 604 observations of a nine-dimensional random vector, on which we applied our method.

We first tested if the obtained variables have linear causal relations. If the variables have linear acyclic causal relations and at most one of the disturbances is Gaussian, the de-mixing matrix obtained by ICA can be permuted to lower-triangularity, and the causal model can be uniquely found [18]. We applied FastICA [7], a widely-used linear ICA algorithm, to the data, and found that the de-mixing matrix is far from a permuted lower-triangular matrix. Furthermore, some independent components are not even truly independent, as verified by the kernel-based independence test [5], which gave the p-value 7×10^{-3}. These findings imply that a linear acyclic causal model does not fit the data well.

We then applied the proposed approach to do nonlinear causal analysis. The results, including the moral graph and the final causal diagram, are shown in Fig. 5. Note that there is a bidirected edge $x_2 \leftrightarrow x_3$ in the final result (Fig. 5(b)); in fact, neither of the causal relations $x_2 \to x_3$ and $x_2 \leftarrow x_3$ could make the disturbance independent of the parents. This means that the causal relation between x_2 and x_3 could not be represented by Eq. 1, or that there exists some confounder. For comparison, the result by the PC algorithm is given in Fig. 5(c). It contains clearly more edges, with two bidirected and three undirected.



(a) moral graph (b) causal diagram (c) by PC

Fig. 5. Experimental results of the MEG data. (a) The moral graph obtained by conditional independence tests. Here no spouse link exists. (b) The final causal model. The thickness of the lines indicates the strength of the causal effects, as measured by the contributed variance. (c) For comparison, the result by the PC algorithm [19] is given.

For illustration, Fig.6 plots the effect of the causes for each variable which has parents. One can see that three of the causal relations ($x_9 \to x_6$, $x_5 \to x_8$, and $(x_1, x_7) \to x_9$) are close to linear, while others are clearly nonlinear. The obtained causal connections are something completely new in neuroscience. Their interpretation will require a lot of work from domain experts.

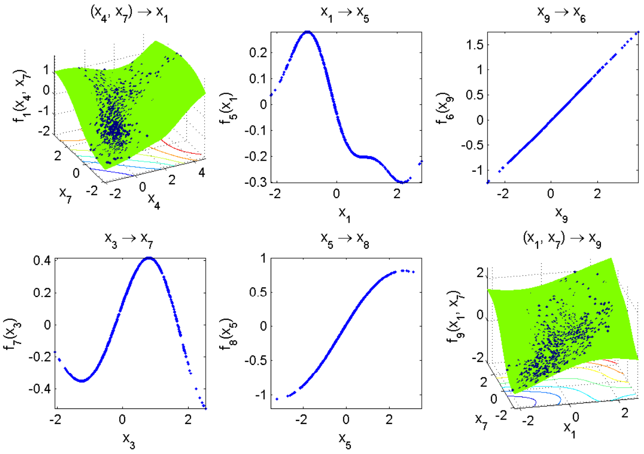


Fig. 6. The effect of the causes on each target variable in the MEG data. For clarity, we also give the fitted surface in 3D plots.

5 Conclusion

In this paper we focused on the acyclic causality discovery problem with an additive disturbance model. For the acyclic causal models in which the causal relations among observed variables are nonlinear while disturbances have linear effects, we have shown that the following criteria are equivalent: 1. mutual independence of the disturbances, 2. causal Markov property of the causal model, as well as the independence between the disturbance and the parents associated with the same variable, and 3. minimum disturbance entropy. From this viewpoint, conventional conditional independence-based methods, non-Gaussianity-based linear methods, and the Granger causality analysis could be unified.

The criterion of mutual independence of disturbances then inspires us to exploit multichannel blind deconvolution, a well-developed extension of ICA, to estimate Granger causality with instantaneous effects. Compared to other methods, this approach is more efficient (in the statistical sense), and it admits simple ways for model selection of the causal structure by incorporating suitable penalties on the coefficients. Finally, we showed that nonlinear causal discovery with additive disturbances can be achieved by enforcing the causal Markov condition and the independence between the disturbance and parents of the same variable.

The resulting approach is suitable for moderate-sized problems. Simulations and real-world applications showed the usefulness of the proposed approaches.

Acknowledgement

We are very grateful to Pavan Ramkumar and Lauri Parkkonen for providing the MEG data. We thank Patrik Hoyer for helpful discussions.

References

1. A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, UK, corrected and revisited edition, 2003.
2. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.
3. P. Gibbs. *Event-Symmetric Space-Time*. Weburbia Press, Great Britain, 1998.
4. C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 1980.
5. 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. MIT Press.
6. 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.
7. A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
8. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.
9. A. Hyvärinen, P. Ramkumar, L. Parkkonen, and R. Hari. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. 2008. Submitted manuscript.
10. A. Hyvärinen, S. Shimizu, and P. O. Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *ICML2008*, pages 424–431, Helsinki, Finland, 2008.
11. R. W. Liu and H. Luo. Direct blind separation of independent non-Gaussian signals with dynamic channels. In *Proc. Fifth IEEE Workshop on Cellular Neural Networks and their Applications*, pages 34–38, London, England, April 1998.
12. D. Margaritis. Distribution-free learning of bayesian network structure in continuous domains. In *Proceedings of the 20th Conference on Artificial Intelligence (AAAI 2005)*, pages 825–830, Pittsburgh, PA, July 2005.
13. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.
14. J.P. Pellet and A. Elisseeff. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 2008.
15. C.E. Rasmussen and C.K.I.Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, USA, 2006.
16. M. Reale and G. Tunnicliffe Wilson. Identification of vector ar models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, 10(1-3):49–65, 2001.

17. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461C464, 1978.
18. 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.
19. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.
20. K. Zhang, H. Peng, L. Chan, and A. Hyvärinen. ICA with sparse connections: Revisisted. In *Proc. 8rd Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2009)*, pages 195–202, Paraty, Brazil, 2009.

Appendix: Proof of Theorem 1

Implication $(iii) \Rightarrow (i)$ is obviously, as shown below. Suppose x_k is neither x_i nor any descendant of x_i. Apparently x_k is a function of the disturbances which do not include e_i, and hence x_k (as well as any parent of x_i) is independent of e_i. According to Eq. 1, we have $p(x_i|pa_i, x_k) = p(e_i|pa_i, x_k) = p(e_i|pa_i) = p(x_i|pa_i)$, meaning x_i and x_k are conditionally independent given pa_i. Below we shall prove $(i) \Rightarrow (ii)$ and $(ii) \Rightarrow (iii)$.

As the causal relations of x_i are acyclic, x_i can be arranged in an order such that no later variable causes any earlier one. Let $k(i)$ denote such an order. Let $\tilde{\mathbf{x}}_i$ be the vector of $\mathbf{x}$ with the order $k(i)$, i.e., $\tilde{x}_{k(i)} = x_i$, and denote by $\tilde{e}_i$ and $\tilde{pa}_i$ the disturbance in $\tilde{x}_i$ and the parents of $\tilde{x}_i$, respectively.

According to the properties of conditional entropy, for any $1 \leq i \leq n$, we have

$$\begin{aligned} H(\tilde{e}_i) &\geq H(\tilde{e}_i|\tilde{pa}_i) && (7) \\ &= H(\tilde{x}_i|\tilde{pa}_i) \geq H(\tilde{x}_i|\tilde{x}_1, \cdots, \tilde{x}_{i-1}), && (8) \end{aligned}$$

where the equality in 7 holds if and only if e_i is independent of $\tilde{pa}_i$, and the equality in 8 holds if and only if $\tilde{x}_j$ ($j < i$, and $\tilde{x}_j \notin \tilde{pa}_i$) are conditionally independent of $\tilde{e}_i$ given $\tilde{pa}_i$ [2]. Summation of the above inequality over all i yields

$$\begin{aligned} \sum_i H(e_i) = \sum_i H(\tilde{e}_i) &\geq H(\tilde{x}_1) + H(\tilde{x}_2|\tilde{x}_1) + \cdots + H(\tilde{x}_n|\tilde{x}_1, \cdots, \tilde{x}_{n-1}) && (9) \\ &= H(\tilde{x}_1, \cdots, \tilde{x}_n) = H(x_1, \cdots, x_n), \end{aligned}$$

where the equality in Eq. 9 holds when (i) is true. This implies $(i) \Rightarrow (ii)$.

Now let us suppose (ii) is true. Denote by $\mathcal{G}$ the transformation from $(\tilde{x}_1, \cdots, \tilde{x}_n)$ to $(\tilde{e}_1, \cdots, \tilde{e}_n)$. As $\tilde{e}_i$ only depends on $\tilde{x}_i$ and its parents, the Jacobian matrix of $\mathcal{G}$, denoted by $\mathbf{J}_\mathcal{G}$, is a lower-triangular matrix with 1 on its diagonal. This gives $|\mathbf{J}_\mathcal{G}| = 1$, and hence $H(\tilde{e}_1, \cdots, \tilde{e}_n) = -E\{\log p_{\tilde{\mathbf{e}}}(\tilde{e}_1, \cdots, \tilde{e}_n)\} = -E\{\log[p_{\tilde{\mathbf{x}}}(\tilde{x}_1, \cdots, \tilde{x}_n)/|\mathbf{J}_\mathcal{G}|]\} = H(x_1, \cdots, x_n)$. Consequently, the mutual information $I(e_1, \cdots, e_n) = \sum_i H(e_i) - H(\tilde{e}_1, \cdots, \tilde{e}_n) = H(x_1, \cdots, x_n) - H(x_1, \cdots, x_n) = 0$, meaning that e_i are mutually independent. We then have $(ii) \Rightarrow (iii)$. Therefore, (i), (ii), and (iii) are equivalent. □