Journal of Machine Learning Research 11 (2010) 1709-1731 Submitted 1/10; Published 5/10

Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Department of Mathematics and Statistics[*]
University of Helsinki
Helsinki, Finland

Kun Zhang KUN.ZHANG@TUEBINGEN.MPG.DE
Department of Computer Science and HIIT
University of Helsinki
Helsinki, Finland

Shohei Shimizu SSHIMIZU@AR.SANKEN.OSAKA-U.AC.JP
Institute of Scientific and Industrial Research
Osaka University
Osaka, Japan

Patrik O. Hoyer PATRIK.HOYER@HELSINKI.FI
Department of Computer Science and HIIT
University of Helsinki
Helsinki, Finland

Editor: Peter Dayan

Abstract

Analysis of causal effects between continuous-valued variables typically uses either autoregressive models or structural equation models with instantaneous effects. Estimation of Gaussian, linear structural equation models poses serious identifiability problems, which is why it was recently proposed to use non-Gaussian models. Here, we show how to combine the non-Gaussian instantaneous model with autoregressive models. This is effectively what is called a structural vector autoregression (SVAR) model, and thus our work contributes to the long-standing problem of how to estimate SVAR's. We show that such a non-Gaussian model is identifiable without prior knowledge of network structure. We propose computationally efficient methods for estimating the model, as well as methods to assess the significance of the causal influences. The model is successfully applied on financial and brain imaging data.

Keywords: structural vector autoregression, structural equation models, independent component analysis, non-Gaussianity, causality

1. Introduction

Analysis of causal influences or effects has become an important topic in statistics and machine learning, and has recently found applications in, for example, neuroinformatics (Roebroeck et al., 2005; Kim et al., 2007) and bioinformatics (Opgen-Rhein and Strimmer, 2007). While the deeper meaning of causality has been formalized in different ways (Pearl, 2000; Spirtes et al., 1993), we

[*] Also in Department of Computer Science and HIIT, University of Helsinki.

consider the problem here from a practical viewpoint, where coefficients in conventional statistical models are interpreted as causal influences.

For continuous-valued variables, such an analysis is typically performed in two different ways. First, if the time-resolution of the measurements is higher than the time-scale of causal influences, one can estimate a classic autoregressive (AR) model with time-lagged variables and interpret the autoregressive coefficients as causal effects. Second, if the measurements have a lower time resolution than the causal influences, or if the data has no temporal structure at all, one can use a model in which the influences are instantaneous, leading to Bayesian networks or structural equation models (SEM); see Bollen (1989).[1]

While estimation of autoregressive methods can be solved by classic regression methods, the case of instantaneous effects is much more difficult. Most methods suffer from lack of identifiability, because covariance information alone is not sufficient to uniquely characterize the model parameters. Prior knowledge of the structure (fixing some of the connections to zero) of the SEM is then necessary for most practical applications. However, a method was recently proposed which uses the non-Gaussian structure of the data to overcome the identifiability problem (Shimizu et al., 2006). If the disturbance variables (external influences) are non-Gaussian, no prior knowledge on the network structure is needed to estimate the linear SEM, except for the ubiquitous assumption of a directed acyclic graph (DAG) and the assumption of no latent variables. (The case of latent variables, that is, unobserved confounders, was later considered by Hoyer et al., 2008.)

Here, we consider the general case where causal influences can occur either instantaneously or with considerable time lags. Such models are one example of structural vector autoregressive (SVAR) models popular in econometric theory, in which numerous attempts have been made for its estimation, see, for example, Swanson and Granger (1997), Demiralp and Hoover (2003) and Moneta and Spirtes (2006). We propose to use non-Gaussianity to estimate the model. We show that this variant of the model is identifiable without other restrictions on the network structure than acyclicity and no latent variables. To our knowledge, no model proposed for this problem has been shown to be fully identifiable without prior knowledge of network structure. We further propose two computationally efficient methods for estimating the model based on the theory of independent component analysis or ICA (Hyvärinen et al., 2001).

The proposed non-Gaussian model not only allows estimation of both instantaneous and lagged effects; it also shows that taking instantaneous influences into account can change the values of the time-lagged coefficients quite drastically. Thus, we see that neglecting instantaneous influences can lead to misleading interpretations of causal effects. The framework further points towards generalizations of the well-known Granger causality measure (Granger, 1969).

The paper is structured as follows. We first define the model and discuss its relation to other models in Section 2. We motivate the key assumption of non-Gaussianity in Section 3. Next, we derive the likelihood and discuss some of its interpretations in Section 4. In Section 5 we propose two computationally efficient estimation methods and compare them with simulations. Assessement of the results using testing is considered in Section 6. Section 7 discusses some interesting phenomena concerning the interpretation of the estimated parameter values. Experiments on financial and neuroscientific data are made in Section 8. Some extensions of the model are discussed in Section 9, and Section 10 concludes the paper. Preliminary results were published in Hyvärinen et al. (2008) and Zhang and Hyvärinen (2009).

[1] Here, we assume that the learning is unsupervised, that is, the inputs to the system are not known or used. If the inputs to the system are known, methods such as dynamic causal modelling can be used (Friston et al., 2003).

2. A Non-Gaussian Structural Vector Autoregressive Model

In this section, we define our new model.

2.1 Background and Notation

Let us denote the observed time series by $x_i(t), i = 1, \ldots, n, t = 1, \ldots, T$ where i is the index of the time series and t is the time index. All the time series (variables) are collected into a single column vector $\mathbf{x}(t)$. Without loss of generality, we can assume that the $x_i(t)$ have zero mean.

In autoregressive modelling, we would model the dynamics by a model of the form

$$\mathbf{x}(t) = \sum_{\tau=1}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t) \tag{1}$$

where k is the number of time-delays used, that is, the order of the autoregressive model, $\mathbf{B}_\tau, \tau = 1, \ldots, k$ are $n \times n$ matrices, and $\mathbf{e}(t)$ is the innovation process.

In structural equation models (SEM), or continuous-valued Bayesian networks, there is no time structure in the data, and the variables are simply modelled as functions of the other variables:

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

where the vector $\mathbf{e}$ is the vector of disturbances or external influences. The diagonal of $\mathbf{B}$ is defined to be zero. It is typically assumed that we have a sample of observations which are independent and identically distributed.

2.2 Definition of Our Model

In many applications, the influences between the $x_i(t)$ can be both instantaneous and lagged. Thus, we combine the two models in (1) and (2) into a single model. Denote by $\mathbf{B}_\tau$ the $n \times n$ matrix of the causal effects between the variables x_i with time lag $\tau, \tau = 0, \ldots, k$. For $\tau > 0$, the effects are ordinary autoregressive effects from the past to the present, while for $\tau = 0$, the effects are "instantaneous".

We define our model by a straightforward combination of (1) and (2) as

$$\mathbf{x}(t) = \sum_{\tau=0}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t) \tag{3}$$

where the $e_i(t)$ are random processes modelling the disturbances. We make the following assumptions on the $e_i(t)$:

1. The $e_i(t)$ are are mutually independent, both of each other and over time. This is a typical assumption in autoregressive models.

2. The $e_i(t)$ are *non-Gaussian*, which is an important assumption which distinguishes our model from classic models, whether autoregressive models, structural-equation models, or Bayesian networks.

3. The matrix modelling instantaneous effects, $\mathbf{B}_0$, corresponds to an *acyclic* graph, as is typical in causal analysis. However this assumption may not be strictly necessary as will be discussed

in Section 9. The acyclicity is equivalent to the existence of a permutation matrix $\mathbf{P}$, which corresponds to a "causal" ordering of the variables x_i, such that the matrix $\mathbf{P}\mathbf{B_0}\mathbf{P}^T$ is lower-triangular (i.e., entries above the diagonal are zero). Acyclicity also implies that the entries on the diagonal are zero, even before such a permutation.

2.3 Relation to Other Models

Next, we discuss the relationships of our model with other models.

2.3.1 Linear Non-Gaussian Acyclic Model

Our model is a generalization of the linear non-Gaussian acyclic model (LiNGAM) proposed in Shimizu et al. (2006). If the order of the autoregressive part is zero, that is, $k = 0$, the model is nothing else than the LiNGAM model, modelling instantaneous effects only. As shown in Shimizu et al. (2006), the assumption of non-Gaussianity of the e_i enables estimation of the model. This is because the non-Gaussian structure of the data provides information not contained in the covariance matrix which is the only source of information in most methods.

Non-Gaussianity enables model estimation using independent component analysis, which solves the non-identifiability of factor analytic models using the assumption of non-Gaussianity of the factors (Comon, 1994; Hyvärinen et al., 2001). In fact, the estimation method in Shimizu et al. (2006) uses an ICA algorithm as an essential part. This is because we can transform (2) into the factor-analytic model of ICA:

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e} \tag{4}$$

where $\mathbf{e}$ is now a vector of latent variables. Under the assumptions of the model, in particular the independence and non-Gaussianity of the disturbances e_i, the model can be essentially estimated (Comon, 1994). The acyclicity assumption also ensures that $\mathbf{I} - \mathbf{B}$ is invertible.

However, there is an important indeterminacy which ICA cannot solve: the order of the components. In a SEM, each disturbance corresponds to one of the observed variables. In contrast, ICA, like most factor-analytic models, gives the factors in no particular order. Thus, after ICA estimation (or as part of the ICA estimation) we have to establish the correspondence between the x_i and the e_i. It was proven by Shimizu et al. (2006) that the correspondence can in fact be established based on the acyclicity of $\mathbf{B}$. Basically, only one of the possible orderings of the rows of $(\mathbf{I} - \mathbf{B})$ is such that all the elements on the diagonal are non-zero, and can thus be scaled to equal one, which has to be the case by definition.

Thus, the LiNGAM model can be estimated by basically first performing ICA estimation and then finding the right ordering of the components based on acyclicity.

2.3.2 Autoregressive Models

On the other hand, if the matrix $\mathbf{B}_0$ has all zero entries, the model in Eq. (3) is a classic vector autoregressive model in which future observations are linearly predicted from preceding ones. If we knew in advance that $\mathbf{B}_0$ is zero, the model could thus be estimated by classic regression techniques since we do not have the same variables on the left and right-hand sides of Eq. (3). However, our model would still be different from classic autoregressive models because the disturbances $e_i(t)$ are non-Gaussian.

It is important to note here that an autoregressive model can serve two different goals: prediction and analysis of causality. Our goal here is the latter: We estimate the parameter matrices $\mathbf{B}_\tau$ in order to interpret them as causal effects between the variables. This goal is distinct from simply predicting future outcomes when passively observing the time series, as has been extensively discussed in the literature on causality (Pearl, 2000; Spirtes et al., 1993). Thus, we emphasize that our model is not intended to reduce prediction errors if we want to predict $x_i(t)$ using (passively) observed values of the past $\mathbf{x}(t-1), \mathbf{x}(t-2), \ldots$; for such prediction, an ordinary autoregressive model is likely to be just as good.

2.3.3 Structural Vector Autoregressive Models

Combinations of SEM and vector autoregressive models have been proposed in the econometric literature, and called structural vector autoregressive models (SVAR). Although many of them are quite similar to our model in spirit (Swanson and Granger, 1997; Demiralp and Hoover, 2003; Moneta and Spirtes, 2006), we are not aware of any method in which non-Gaussianity would be an essential assumption. We will see below how the assumption of non-Gaussianity is essential for the identifiability of the model, which has been a serious problem in previous methods. In the next section, we thus consider the justification of this assumption.

3. Why Disturbances Could be Non-Gaussian

Non-Gaussianity is the new assumption in our model. In this section, we attempt to justify why, in many applications, we can consider the $e_i(t)$ to be non-Gaussian. The arguments are based on heteroscedasticity. We do not by any means claim that we are the first to develop these arguments; some of them are well-known and we merely re-iterate them here.

The principle of heteroscedasticity means that the variance of $e_i(t)$ depends on t: in some parts of the time series, it is larger, and in other parts it is smaller. The shape of the distribution conditional to the variance is the same always: often it is assumed to be Gaussian (normal).

We argue that heteroscedasticity is an important reason why, in many cases, the $e_i(t)$ should be strongly non-Gaussian. Even if the Central Limit Theorem is applicable in the sense that $e_i(t)$ is a sum of many different latent independent variables, the disturbances can be very non-Gaussian if, for some reason, the variance of the $e_i(t)$ is changing.

The connection between heteroscedasticity and non-Gaussianity can be developed in a few simple equations. Denote by $z(t)$ a standardized Gaussian random variable. Assume that a disturbance $e(t)$ (dropping the index i for simplicity) is a product of z and a random "variance" variable $v(t)$:

$$e(t) = z(t)v(t)$$

where $z(t)$ and $v(t)$ are independent by definition. We can, in fact, drop the time indices and just consider these time series as random variables. The distribution of v can be of different kinds, whereas the distribution of z is fixed to standardized Gaussian. In the simplest case, v takes only two different values, which means that the data points belong to just two different classes, and the density is then a finite mixture of two Gaussian distributions.

We can simply show the following well-known result: If z is Gaussian, e has always positive kurtosis,[2] regardless of the distribution of v (as long as v^2 has non-zero variance). This is because

$$\text{kurt}(e) = E\{e^4\} - 3(E\{e^2\})^2 = E\{v^4 z^4\} - 3(E\{v^2 z^2\})^2 = 3[E\{v^4\} - (E\{v^2\})^2] \tag{5}$$

which is always positive because it is the variance of v^2 multiplied by three (Beale and Mallows, 1959). It is easy to generalize this result to show that even if z is not Gaussian, the kurtosis is still positive if the variance of v^2 is large enough.

Heteroscedastity can be seen in some important application areas of causal modelling, in particular:

1. In econometrics, heteroscedastic models have a long tradition (Engle, 1995). For example, in financial markets the volatility of a price is often assumed to be changing over time, and volatility is nothing but the variance in some scaling.

2. In brain imaging, the power of rhytmic activity as measured by electroencephalography or magnetoencephalography is non-constant (Hari and Salmelin, 1997). The power is essentially the same as the variance.

We emphasize that the assumption of non-Gaussianity is fundamentally an empirical assumption. It is fulfilled in some application areas and not in others. It can be validated by examining the distributions of the estimates of the $e_i(t)$, which are simply obtained by solving for $\mathbf{e}(t)$ in (3) after estimation of the model. Those estimates are linear functions of the data, which implies that if the data were Gaussian, the $e_i(t)$ would necessarily be Gaussian. Thus, any non-Gaussianity in the estimates is valid evidence for the Gaussianity of the underlying $e_i(t)$. In addition to visual inspection, any formal tests for non-Gaussianity can be used, such as the Shapiro-Wilk test or the Kolmogorov-Smirnov test. (Independence of the $e_i(t)$ can be validated in the same way, although it seems to be more difficult to investigate by visualization or testing.)

However, in practice the question is not whether the disturbances are non-Gaussian but whether they are sufficiently far from Gaussian to enable sufficiently accurate estimation. In the theory of ICA, it has been shown that the asymptotic variance of the estimators is a function of the non-Gaussianity of the components: When their distribution approaches Gaussianity, the asymptotic variance goes to infinity (Cardoso and Laheld, 1996; Hyvärinen et al., 2001). Thus, instead of testing non-Gaussianity it may be much more useful to simply measure the accuracy of the estimates by bootstrapping and similar methods. If the disturbances are Gaussian (or very close to Gaussian), our estimation method is likely to fail completely. Some other assumptions are then needed to obtain identifiability of the model.

It should be also noted that the assumptions of non-Gaussianity and independence cannot be easily disentangled from the assumption of linearity. If there are non-linearities in the system, these may, for example, lead to non-Gaussian residuals even if the original disturbances were Gaussian.

4. Likelihood of the Model

To estimate our model, we start by formulating its likelihood.

[2] We use here the definition of kurtosis given in Eq. (5), which is sometimes called excess kurtosis. Thus, kurtosis can be either positive or negative.

4.1 Likelihood of LiNGAM

First, we derive the likelihood of the LiNGAM model (Shimizu et al., 2006) which has not yet been given in the literature. The starting point is the likelihood of the ICA model which is well-known, see, for example, Pham and Garrat (1997) and Hyvärinen et al. (2001). Denote the ICA model by

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

for a square invertible matrix $\mathbf{A}$, and independent non-Gaussian latent variables s_i. Denote the observed sample by $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$ and $\mathbf{W} = \mathbf{A}^{-1}$. The log-likelihood is then usually given in the form

$$\log L_0(\mathbf{X}) = \sum_t \sum_i \log p_i(\mathbf{w}_i^T \mathbf{x}(t)) + T \log \det |\mathbf{W}|$$

where the p_i are the density functions of the independent components (here: disturbances). Since the densities of the disturbances are not specified, we have in general a semi-parametric problem. Different methods have been developed for approximating $\log p_i$, for example, Pham and Garrat (1997), Karvanen and Koivunen (2002) and Chen and Bickel (2006). Here, we have to take into account the fact that those methods usually assume that the independent components have been normalized to unit variance, which is not the case in LiNGAM. Thus, we prefer to modify the formula by normalizing the densities as follows:

$$\log L_1(\mathbf{X}) = \sum_t \sum_i \log \tilde{p}_i(\frac{\mathbf{w}_i^T \mathbf{x}(t)}{\sigma_i}) - T \sum_i \log \sigma_i + T \log \det |\mathbf{W}| \qquad (6)$$

where the $\tilde{p}_i$ are the densities of the disturbances standardized to unit variance, and the σ_i^2 are their variances before standardization.

In fact, in practice it has been realized that often one can just fix the $\tilde{p}_i$ to a single function and still obtain a satisfactory estimator. In particular, if we know that the disturbances are all super-Gaussian (i.e., have positive kurtosis), fixing

$$\log \tilde{p}_i(s) = -\sqrt{2}|s| + \text{const.}$$

is enough to provide a consistent estimator under weak constraints (Cardoso and Laheld, 1996; Hyvärinen and Oja, 1998).

In LiNGAM, we have from (4) that in terms of the ICA model, $\mathbf{A}^{-1} = \mathbf{W} = \mathbf{I} - \mathbf{B}_0$ (we use the subindex 0 for $\mathbf{B}$ in LiNGAM to comply with the notation below). Now, we can simplify the likelihood because of the DAG structure. The DAG structure means that for the right permutation of its rows (corresponding to the causal ordering), $\mathbf{W}$ is lower-triangular. The determinant of a triangular matrix is equal to the product of its diagonal elements, and a permutation does not change the determinant, so the determinant of $\mathbf{W}$ is equal to the product of the diagonal elements when the variables are ordered in the causal order. But by definition of $\mathbf{W}$ in LiNGAM, those diagonal elements are all equal to one, so the last term in (6) is zero. So, the likelihood of the LiNGAM model is finally given by

$$\begin{aligned}\log L(\mathbf{X}) &= \sum_t \sum_i \log \tilde{p}_i \left(\frac{\mathbf{w}_i^T \mathbf{x}(t)}{\sigma_i} \right) - T \sum_i \log \sigma_i \\ &= \sum_t \sum_i \log \tilde{p}_i \left(\frac{x_i(t) - \mathbf{b}_{0,i}^T \mathbf{x}(t)}{\sigma_i} \right) - T \sum_i \log \sigma_i \qquad (7)\end{aligned}$$

where the variances of the components can be estimated by taking the empirical variances as

$$\sigma_i^2 = \frac{1}{T}\sum_t (x_i(t) - \mathbf{b}_{0,i}^T \mathbf{x}(t))^2.$$

(Alternatively, the σ_i could be obtained by a separate maximization of the likelihood, but this would be more complicated computationally and conceptually.) Here, $\mathbf{W}$ is constrained to correspond to a DAG, with ones added in the diagonal.

4.2 Likelihood of Our Model

Now we can derive the likelihood of our model. First note that from (3) we can derive

$$\mathbf{x}(t) = \sum_{\tau=0}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t) \Leftrightarrow \ (\mathbf{I} - \mathbf{B}_0)[\mathbf{x}(t) - \sum_{\tau=1}^{k} (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_\tau \mathbf{x}(t-\tau)] = \mathbf{e}(t),$$

which gives

$$\mathbf{x}(t) - \sum_{\tau=1}^{k} (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_\tau \mathbf{x}(t-\tau) = \mathbf{B}_0[\mathbf{x}(t) - \sum_{\tau=1}^{k} (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_\tau \mathbf{x}(t-\tau)] + \mathbf{e}(t)$$

which shows that the our model in (3) is a LiNGAM model for the residuals $\mathbf{x}(t) - \sum_{\tau=1}^{k}(\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_\tau \mathbf{x}(t-\tau)$. Denoting

$$\mathbf{M}_\tau = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_\tau \text{ and } \mathbf{W} = \mathbf{I} - \mathbf{B}_0 \tag{8}$$

and replacing $\mathbf{x}(t)$ in (7) by the residuals, we have

$$\log L(\mathbf{X}) = \sum_t \sum_i \log \tilde{p}_i \left(\frac{\mathbf{w}_i^T [\mathbf{x}(t) - \sum_{\tau=1}^{k} \mathbf{M}_\tau \mathbf{x}(t-\tau)]}{\sigma_i} \right) - \log \sigma_i \tag{9}$$

with

$$\sigma_i^2 = \frac{1}{T} \sum_t \left(\mathbf{w}_i^T [\mathbf{x}(t) - \sum_{\tau=1}^{k} \mathbf{M}_\tau \mathbf{x}(t-\tau)] \right)^2.$$

4.3 Information-Theoretic Interpretation

An interesting point to note is that the likelihood is now a sum of the negative entropies of the residuals. The differential entropy of a random variable s can be written using the standardized version of s, denoted by $\tilde{s}$, as follows:

$$H(s) = -\int p_s(u) \log p_s(u) du = -\int p_{\tilde{s}}(u) \log p_{\tilde{s}}(u) du + \log \sigma_s$$

where σ_s^2 is the variance of s. Thus, we can interpret the terms in (9) as the (negative) entropies of the residuals. So, estimation is accomplished by minimizing the "prediction errors" or "uncertainties" in the DAG if the entropies are interpreted as the prediction errors when each variable is predicted by its parents. Note that for Gaussian variables, the entropies are very simple functions of the squared errors (variances), while for non-Gaussian variables, they are also functions of the non-Gaussianity (shape) of the distribution.

5. Practical Estimation Methods

Next we propose two practical methods for estimating the model, and further show how to include a sparseness penalty which may be very useful in practice.

5.1 A Two-Stage Method with Least-Squares Estimation

Optimization of the likelihood is difficult because it contains a complicated combinatorial optimization part due to the constraint that $\mathbf{B}_0$ is acyclic. A conceptually simple way of reinforcing this constraint would be to go through all possible orderings of the observed variables, and for each of them, maximize the likelihood as a function of the $\mathbf{B}_\tau$ so that $\mathbf{B}_0$ is constrained to be lower triangular. This is obviously computationally very expensive since the number of ordering is equal to $n!$ where n is the number of variables. Only for a very small n can this be computationally feasible. (Another difficulty is that the likelihood contains a semiparametric part because we do not specify a parametric model of the non-Gaussian distributions, but this problem has already been extensively treated in the theory of ICA, and has been found not to be very serious in practice, see Hyvärinen et al., 2001.)

To avoid the computational problems with likelihood, we propose a computationally simpler two-stage method for estimating our model. The method combines classic least-squares estimation of an autoregressive (AR) model with LiNGAM estimation.

5.1.1 DEFINITION

The basic idea is that the $\mathbf{M}_\tau$ in (8) can be consistently, and computationally efficiently, estimated by classic least-squares methods. Then, since the model is essentially a LiNGAM model for the residuals of the predictions by the $\mathbf{M}_\tau$, we simply use our previously developed estimation methods for LiNGAM to estimate the rest of the parameters. These methods (Shimizu et al., 2006) seem to tackle the combinatorial optimization problem in a satisfactory way. The ensuing method will be justified in more detail below; it is defined as follows:

1. Estimate a classic autoregressive model for the data

$$\mathbf{x}(t) = \sum_{\tau=1}^{k} \mathbf{M}_\tau \mathbf{x}(t-\tau) + \mathbf{n}(t) \tag{10}$$

   using any conventional implementation of a least-squares method. Note that here $\tau > 0$, so it is really a classic AR model. Denote the least-squares estimates of the autoregressive matrices by $\hat{\mathbf{M}}_\tau$.

2. Compute the residuals, that is, estimates of $\mathbf{n}(t)$

$$\hat{\mathbf{n}}(t) = \mathbf{x}(t) - \sum_{\tau=1}^{k} \hat{\mathbf{M}}_\tau \mathbf{x}(t-\tau).$$

3. Perform the LiNGAM analysis (Shimizu et al., 2006) on the residuals. This gives the estimate of the matrix $\mathbf{B}_0$ as the solution of the instantaneous causal model

$$\hat{\mathbf{n}}(t) = \mathbf{B}_0 \hat{\mathbf{n}}(t) + \mathbf{e}(t).$$

4. Finally, compute the estimates of the causal effect matrices $\mathbf{B}_\tau$ for $\tau > 0$ as

$$\hat{\mathbf{B}}_\tau = (\mathbf{I} - \hat{\mathbf{B}}_0)\hat{\mathbf{M}}_\tau \text{ for } \tau > 0. \tag{11}$$

5.1.2 Consistency Proof

We now prove that this provides a consistent estimator of $\mathbf{B}_\tau$. First, the basic model definition in (3) can be manipulated to yield

$$(\mathbf{I} - \mathbf{B}_0)\mathbf{x}(t) = \sum_{\tau=1}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t)$$

and thus

$$\mathbf{x}(t) = \sum_{\tau=1}^{k} (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_\tau \mathbf{x}(t-\tau) + (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{e}(t). \tag{12}$$

Now, a well-known result is that least-squares estimation of an AR model as in (10) is consistent even if the innovation vector $\mathbf{n}(t)$ does not have independent or even uncorrelated elements (for fixed t), and even if it is heteroscedastic and non-Gaussian. Thus, comparing (12) with (10), in the limit we can equate the autoregressive matrices, which gives $(\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_\tau = \mathbf{M}_\tau$ for $\tau \geq 1$, and thus (11) is justified. (In fact, we anticipated (11) in the notation used in the likelihood in (9).)

Second, comparison of (12) with (10) shows that the residuals $\hat{\mathbf{n}}(t)$ are, asymptotically, of the form $(\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{e}(t)$. This means

$$\hat{\mathbf{n}}(t) = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{e}(t) \Leftrightarrow (\mathbf{I} - \mathbf{B}_0)\hat{\mathbf{n}}(t) = \mathbf{e}(t) \Leftrightarrow \hat{\mathbf{n}}(t) = \mathbf{B}_0\hat{\mathbf{n}}(t) + \mathbf{e}(t)$$

which is the LiNGAM model for $\hat{\mathbf{n}}(t)$. This shows that $\mathbf{B}_0$ is obtained as the LiNGAM analysis of the residuals, and the consistency of our estimator of $\mathbf{B}_0$ follows from the consistency of LiNGAM estimation (Shimizu et al., 2006). Thus, our method is consistent for all the $\mathbf{B}_\tau$. This obviously proves, by construction, the identifiability of the model as well.

5.1.3 Interpretation Related to ICA of Residuals

An interesting viewpoint of the two-stage estimation method is analysis of the dependencies of the innovations after estimating a classic AR model. Suppose we just estimate an AR model as in (1), and interpret the coefficients as causal effects. Such an interpretation more or less presupposes that the innovations $e_i(t)$ are independent of each other, because otherwise there is some structure in the model which has not been modelled by the AR model. If the innovations are not independent, the causal interpretation may not be justified. Thus, it seems necessary to further analyze the dependencies in the innovations in cases where they are strongly dependent.

Analysis of the dependency structure in the residuals (which are, by definition, estimates of innovations) is precisely what leads to the two-stage estimation method. As a first approach, one could consider application of something like principal component analysis or independent component analysis on the residuals. The problem with such an approach is that the interpretation of the obtained results in the framework of causal analysis would be quite difficult. Our solution is to fit a causal model like LiNGAM to the residuals, which leads to a straightforward causal interpretation of the analysis of residuals which is logically consistent with the AR model.

5.2 Method Based on Multichannel Blind Deconvolution

While the two-stage method proposed above is computationally very efficient, it is far from being statistically optimal. The estimation of the autoregressive part takes in no way non-Gaussianity into account and is thus likely to be suboptimal. However, it is useful because it provides a good initial guess for any further iterative method.

Thus, to improve the results of the two-stage method, we further propose an estimation method based on the similarity of our model with convolutive versions of ICA which are also called multichannel blind deconvolution (MBD). Estimation of the model Eq. (3) is, in fact, closely related to the multichannel blind deconvolution problem with causal finite impulse response (FIR) filters (Cichocki and Amari, 2002; Hyvärinen et al., 2001). MBD, as a direct extension of ICA, assumes that the observed signals are convolutive mixtures of some spatially and independently and identically distributed (i.i.d.) sources.

Using MBD methods is justified here due to the possibility or transforming an autoregressive model into a moving-average model: In Eq. (3), the observed variables $x_i(t)$ can be considered as convolutive mixtures of the disturbances $e_i(t)$. Thus, we can find estimates of $\mathbf{B}_\tau$, as well as $e_i(t)$, in Eq. (3), by using MBD methods to estimate the filter matrices $\mathbf{W}_\tau$

$$\hat{\mathbf{e}}(t) = \sum_{\tau=0}^{k} \mathbf{W}_\tau \mathbf{x}(t-\tau). \qquad (13)$$

Comparing (13) with (3), we can see that the $\mathbf{B}_\tau$ can then be recovered from the estimated $\mathbf{W}_\tau$; details are given below.

The basic statistical principle to estimate the MBD model is that the disturbances $e_i(t)$ should be mutually independent for different i and different t. 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 $e_i(t)$ and $\mathbf{B}_\tau$) up to some scaling, permutation, and time shift indeterminacies (Liu and Luo, 1998). This implies that our SVAR model is identifiable by MBD if at most one of the disturbances e_i is Gaussian.

There exist several well-developed algorithms for MBD. For example, one may adopt the one based on natural gradient (Cichocki and Amari, 2002). By extending the LiNGAM analysis procedure (Shimizu et al., 2006), 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 $\widetilde{\mathbf{W}}_0$ with only significantly nonzero entries on the main diagonal. The permutation can be found using similar methods (e.g., the Hungarian algorithm) as in LiNGAM (Shimizu et al., 2006). 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 final estimates of $\mathbf{B}_0$ and $\mathbf{B}_\tau$ ($\tau > 0$) can then 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.

5.3 Sparsification of the Causal Connections

For the purposes of interpretability and generalizability, it is often useful to sparsify the causal connections, that is, to set insignificant entries of $\hat{\mathbf{B}}_\tau$ to zero. Analogously to the development of ICA with sparse connections (Zhang et al., 2009), we can incorporate an adaptive L_1 penalty into the likelihood of the MBD method to achieve fast model selection which performs such sparsification. We use a penalty-based approach because traditional model selection based on information criteria involves a combinatorial optimization problem whose complexity increases exponentially in the dimensionality of the parameter space. In the MBD problem, this is often not computationally feasible.

Thus, to make $\mathbf{W}_\tau$ in Eq. (13) as sparse as possible, we maximize the penalized likelihood defined as

$$pl(\{\mathbf{W}_\tau\}) = \log L(\{\mathbf{W}_\tau\}) - \lambda \sum_{i,j,\tau} |w_{i,j,\tau}|/|\hat{w}_{i,j,\tau}|, \tag{14}$$

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. The parameter λ is the general weight of the penalty.

The idea here is that we first compute an initial estimate of the $w_{i,j,\tau}$ by a conventional method (such as maximum likelihood) and then use those estimates to compute a parameter-wise weighting in the L_1 penalty. The end result is that those $w_{i,j,\tau}$ for which the initial estimates $\hat{w}_{i,j,\tau}$ were small are heavily penalized, and likely to be zero in the final estimate obtained by maximization of pl.

This penalized likelihood is a special case of adaptive Lasso and therefore has the same consistency in variable selection (Zou, 2006). In fact, it can also be used for selecting the order k of the autoregressive model. In particular, to achieve model selection similar to the Bayesian Information Criterion (BIC), one can set $\lambda = \log T$, where T is the sample size (Zhang et al., 2009).

It may be also useful to penalize groups of parameters. In particular, to see if the historical values of $x_i(t)$ causes $x_j(t)$ $(i \neq j)$, one needs to examine the combined effect of the group of parameters $[\hat{\mathbf{B}}_\tau]_{i,j}, \tau = 1,...,p$, and therefore it makes sense to apply penalization on the parameter group. Combining the above approach with group Lasso (Bach, 2008) leads to the following penalized likelihood:[3]

$$pl(\{\mathbf{W}_\tau\}) = \log L(\{\mathbf{W}_\tau\}) - \lambda \sum_{i,j,\tau} |w_{i,j,0}|/|\hat{w}_{i,j,0}| - k\lambda \sum_{i,j} \Big(\sum_{\tau=1}^{k} w_{i,j,\tau}^2\Big)^{1/2} \Big/ \Big(\sum_{\tau=1}^{k} \hat{w}_{i,j,\tau}^2\Big)^{1/2},$$

where the last term has the coefficient k because the parameter group $w_{i,j,\tau}, \tau = 1,...,k$ has k parameters.

5.4 Simulations

To investigate the performance of the proposed estimation methods, we conducted a series of simulations. We set the number of lags $k = 1$ and the dimensionality $n = 5$. We randomly constructed the strictly lower-triangular matrix $\mathbf{B}_0$ and matrix $\mathbf{B}_1$. To make the causal effects sparse, we set about 60% of the entries in the matrix $\mathbf{B}_1$ and the lower-triangular part of $\mathbf{B}_0$ to zero, while the magnitude of the others is uniformly distributed between 0.05 and 0.5 and the sign is random. Super-Gaussian

[3] Here we treat the instantaneous effects separately. If one would like to see if the total influence from $x_i(t-\tau), \tau = 0,1,...,p$ to $x_j(t)$ is significant, all parameters $w_{i,j,\tau}, \tau = 0,1,...,p$ should be treated as a group.

disturbances $e_i(t)$ were generated by passing standardized i.i.d. Gaussian variables through a power nonlinearity with exponent between 1.5 and 2.0 while keeping the original sign. The observations $\mathbf{x}(t)$ were then generated according to the model in Eq. (3). Various sample sizes (T = 100, 300, and 1000) were tested. We compared the performance of the two-stage method (Section 5.1), the method by MBD (Section 5.2) and the MBD-based method with the sparsity penalty (Section 5.3). In the last method, we set the penalization parameter in Eq. (14) as $\lambda = \log T$ to make its results consistent with those obtained by BIC. The densities of the independent components were adaptively estimated using the method in Pham and Garrat (1997). 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-stage 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 penalty performed best.

6. Assessment of the Significance of Causality

In practice, we also need to assess the significance of the estimated causal relations. While the sparsification method in Section 5.3 is related to this goal, here we propose a more principled approach for testing the significance of the causal influences.

For the instantaneous effects $x_i(t) \to x_j(t)$ $(i \neq j)$, the significance of causality is obtained by assessing if the entries of $\hat{\mathbf{B}}_0$ are statistically significantly different from zero. For the lagged effects $x_i(t-\tau) \to x_j(t)$ $(i \neq j, \tau > 0)$, however, one is often not interested in the significance of any single coefficient in $\hat{\mathbf{B}}_\tau$: More frequently one aims to find out if the total effect from $x_i(t-\tau)$ to $x_j(t)$ is significant.

We propose two simple statistics. One is a measure of instantaneous variance contributed by $x_i(t)$ to $x_j(t)$: $S_0(i \leftarrow j) = [\mathbf{B}_0]^2_{ij} \cdot \mathrm{var}(x_i(t))/\mathrm{var}(x_j(t))$. If all time series have the same variance, it is simplified to $S_0(i \leftarrow j) = [\mathbf{B}_0]^2_{ij}$. The other measures how strong the total lagged causal influence from $x_i(t)$ to $x_j(t)$ is; it is a measure of contributed variance from $x_i(t-\tau), \tau > 0$ to $x_j(t)$: $S_{lag}(i \leftarrow j) = \mathrm{var}(\sum_{\tau>0}[\mathbf{B}_\tau]_{ij} x_j(t-\tau))/\mathrm{var}(x_j(t))$. If all series $x_i(t)$ have the same variance and are approximately temporally uncorrelated, the above statistic can be approximated by $\sum_{\tau>0}[\mathbf{B}_\tau]^2_{ij}$. (Note that these quantities are not exactly proportions of variance explained because the explaining variables are not necessarily uncorrelated.)

The asymptotic distributions of these statistics under the null hypothesis (with no causal effects) are very difficult to derive, and they may also behave poorly in the finite sample case. Therefore, like in Diks and DeGoede (2001) and Theiler et al. (1992), we use bootstrapping with surrogate data to find the empirical distributions of each statistic under the null hypothesis. To generate the surrogate data under the null hypothesis, in each bootstrapping replication we “scramble” the original series $x_i(t)$, that is, each time series is randomly permuted in temporal order. We then calculate $\hat{S}^*$, the estimate of the statistic S (which may be $S_0(i \leftarrow j)$ or $S_{lag}(i \leftarrow j)$) for the surrogate data. Next, the α-level bootstrapping critical value $c^*_{t\alpha}$ is found as the α-th quantile of the bootstrapping distribution

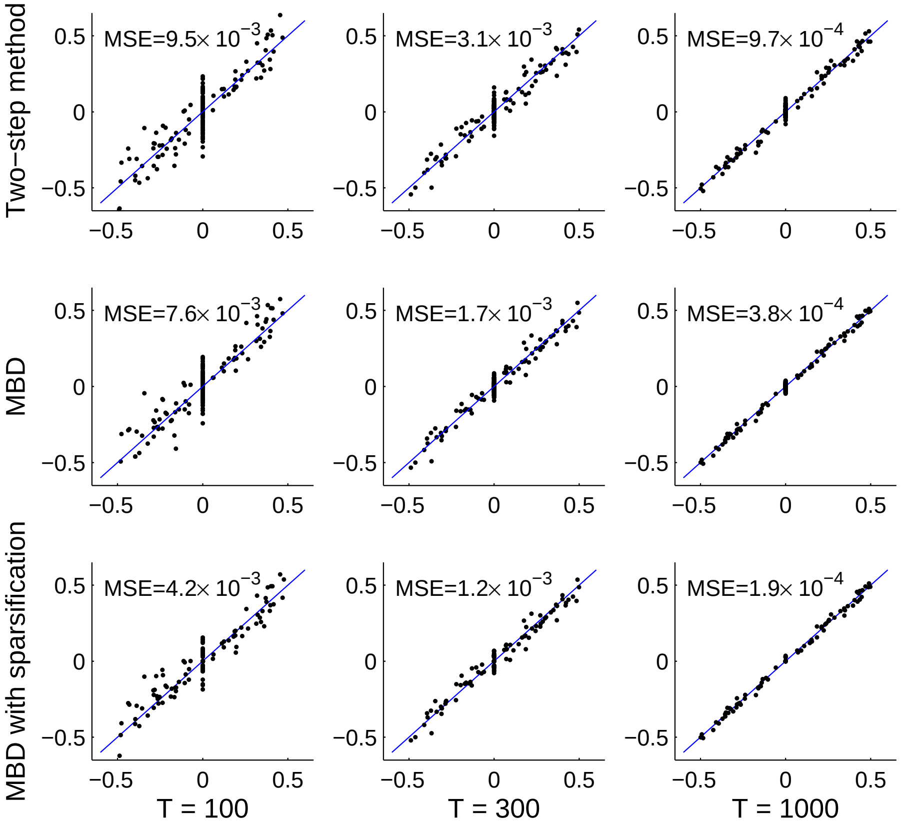


Figure 1: Scatter plots of the estimated coefficients (y axis) versus the true ones (x axis) for different sample sizes and different methods.

of $\hat{S}^*$. Finally, we reject the null hypothesis if $\hat{S} > c^*_{t\alpha}$, where $\hat{S}$ is the estimate of S for the original data.

7. Remarks on the Interpretation of the Parameters

In this section, we discuss how the autoregressive parameters are changed by taking into account the instantaneous effects, and how our model can be interpreted in the framework of Granger causality.

7.1 Interaction Between Instantaneous and Lagged Effects

Equation (11) shows the interesting fact already mentioned in the Introduction: Consistent estimates of the $\mathbf{B}_\tau$ are not obtained by a simple AR model fit even for $\tau > 0$. Taking instantaneous effects into account changes the estimation procedure for all the autoregressive matrices, if we want consistent estimators as we usually do. Of course, this is only the case if there are instantaneous effects, that is, $\mathbf{B}_0 \neq 0$; otherwise, the estimates are not changed.

While this phenomenon is, in principle, well-known in econometric literature (Swanson and Granger, 1997; Demiralp and Hoover, 2003; Moneta and Spirtes, 2006), Eq. (11) is seldom applied because estimation methods for $\mathbf{B}_0$ have not been well developed. To our knowledge, no estimation method for $\mathbf{B}_0$ has been proposed which is consistent for the whole matrix without strong prior assumptions on $\mathbf{B}_0$.

Next we present some theoretical examples of how the instantaneous and lagged effects interact based on the formula in (11).

7.1.1 EXAMPLE 1: AN INSTANTANEOUS EFFECT MAY SEEM TO BE LAGGED

Consider first the case where the instantaneous and lagged matrices are as follows:

$$\mathbf{B}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix}.$$

That is, there is an instantaneous effect $x_2 \to x_1$, and no lagged effects (other than the purely autoregressive $x_i(t-1) \to x_i(t)$). Now, if an AR(1) model is estimated for data coming from this model, without taking the instantaneous effects into account, we get the autoregressive matrix

$$\mathbf{M}_1 = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1 = \begin{pmatrix} 0.9 & 0.9 \\ 0 & 0.9 \end{pmatrix}.$$

Thus, the effect $x_2 \to x_1$ seems to be lagged although it is, actually, instantaneous.

7.1.2 EXAMPLE 2: SPURIOUS EFFECTS APPEAR

Consider three variables with the instantaneous effects $x_1 \to x_2$ and $x_2 \to x_3$, and no lagged effects other than $x_i(t-1) \to x_i(t)$, as given by

$$\mathbf{B}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.9 \end{pmatrix}.$$

If we estimate an AR(1) model for the data coming from this model, we obtain

$$\mathbf{M}_1 = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0.9 & 0.9 & 0 \\ 0.9 & 0.9 & 0.9 \end{pmatrix}.$$

This means that the estimation of the simple autoregressive model leads to the inference of a direct lagged effect $x_1 \to x_3$, although no such direct effect exists in the model generating the data, for any time lag.

A more reassuring result is the following: if the data follows the same causal ordering for all time lags, that ordering is not contradicted by the neglect of instantaneous effect. A rigorous definition of this property is the following.

Theorem 1 *Assume that there is an ordering $i(j), j = 1 \ldots n$ of the variables such that no effect goes backward, that is,*

$$\mathbf{B}_\tau(i(j-\delta), i(j)) = 0 \textit{ for } \delta > 0, \tau \geq 0, 1 \leq j \leq n. \tag{15}$$

(In the purely instantaneous case, existence of such an ordering is equivalent to acyclicity of the effects as noted in Section 2.2.) Then, the same property applies to the $\mathbf{M}_\tau, \tau \geq 1$ *as well. Conversely, if there is an ordering such that (15) applies to* $\mathbf{M}_\tau, \tau \geq 1$ *and* $\mathbf{B}_0$*, then it applies to* $\mathbf{B}_\tau, \tau \geq 1$ *as well.*

Proof: When the variables are ordered in this way (assuming such an order exists), all the matrices $\mathbf{B}_\tau$ are lower-triangular. The same applies to $\mathbf{I} - \mathbf{B}_0$. Now, the product of two lower-triangular matrices is lower-triangular; in particular the $\mathbf{M}_\tau$ are also lower-triangular according to (11), which proves the first part of the theorem. The converse part follows from solving for $\mathbf{B}_\tau$ in (11) and the fact that the inverse of a lower-triangular matrix is lower-triangular.

What this theorem means is that if the variables really follow a single "causal ordering" for all time lags, that ordering is preserved even if instantaneous effects are neglected and a classic AR model is estimated for the data. Thus, there is some limit to how (11) can change the causal interpretation of the results.

7.2 Generalizations of Granger Causality

The classic interpretation of causality in instantaneous SEMs would be that x_i causes x_j if the (j,i)-th coefficient in $\mathbf{B}_0$ is non-zero. On the other hand, in the time series context, the concept of Granger causality (Granger, 1969) formalizes causality as the ability to reduce prediction error. A simple operational definition of Granger causality can be based on the autoregressive coefficients $\mathbf{M}_\tau$: If at least one of the coefficients from $x_i(t-\tau), \tau \geq 1$ to $x_j(t)$ is (significantly) non-zero, then x_i Granger-causes x_j. This is because then the variable x_i reduces the prediction error in x_j in the mean-square sense if it is included in the set of predictors, which is the very definition of Granger causality.

In light of the results in this paper, we can generalize the concept of Granger causality in two ways. First we can combine the two aspects of instantaneous and lagged effects. In fact, such a concept of instantaneous causality was already alluded to by Granger (1969), but presumably due to lack of proper estimation methods, that paper as well as most future developments considered mainly non-instantaneous causality. The second generalization is to measure prediction error by the information-theoretic definition of Section 4.3, essentially using entropy instead of mean squared error. These two generalization are independent of each other in the sense that we can use any one of them, omitting the other.

Both of these extensions are implicit in estimation of our model. Thus, we define that *a variable* x_i *causes* x_j *if at least one of the coefficients* $\mathbf{B}_\tau(j,i)$*, giving the effect from* $x_i(t-\tau)$ *to* $x_j(t)$*, is (significantly) non-zero for* $\tau \geq 0$. The condition for τ is different from Granger causality since the value $\tau = 0$, corresponding to instantaneous effects, is included. Moreover, since estimation of the instantaneous effects changes the estimates of the lagged ones, the lagged effects used in our definition are different from those usually used with Granger causality. Using entropy instead of mean-squared error is implicit in this definition because non-Gaussianity is used in the estimation of the model. In general, entropy minimization is closely related to ICA estimation (Hyvärinen, 1999) as well as the estimation of the present model as was discussed in Section 4.3. Notice that we assume here, as in the general theory of Granger causality, that there are no unobserved confounders.

8. Real Data Experiments

We applied our model together with the estimation and testing method on both financial data and magnetoencephalography (MEG) data. In the former application, we used the sparsity penalty to select significant effects, while in the latter one, bootstrapping was used.

8.1 Application in Finance

First, we use the model in Eq. (3) 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 4th Dec 2001 to 11th Jul 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 $P_i(t)$, the corresponding return is calculated by $x_i(t) = \frac{P_i(t)-P_i(t-1)}{P_i(t-1)}$. The data for analysis are $\mathbf{x}(t) = [x_1(t),...,x_4(t)]^T$, with $T = 1200$ observations.

We applied the MBD-based method with the sparsity penalty to $\mathbf{x}(t)$. The kurtoses of the estimated disturbances $\hat{e}_i$ 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 exactly zero due to sparsity penalty. $\widehat{\mathbf{B}}_0$ and $\widehat{\mathbf{B}}_1$ were constructed based on $\mathbf{W}_0$ and $\mathbf{W}_1$, using the procedure given in Section 5.2. 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, which is 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 Asia and USA. That is, the causal effects from N225_t and HSI_t to DJI_t, although seeming to be instantaneous, may actually be 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.

8.2 Application on MEG Data

Second, we applied the proposed model on the magnitudes of brain sources obtained from magnetoencephalographic (MEG) signals to analyze their causal relationships. The raw recordings consisted of the 204 gradiometer channels measured by the Vectorview helmet-shaped neuromagnetometer (Neuromag Ltd., Helsinki, Finland) in a magnetically shielded room at the Brain Research Unit of the Low Temperature Laboratory of the Aalto University School of Science and Technology. They were obtained from a healthy volunteer and lasted about 12 minutes. The data was downsampled to 75 Hz.

To begin with, we separated sources underlying the recorded MEG data using a recently proposed blind source separation method, Fourier-ICA (Hyvärinen et al., 2010). We manually selected 17 sources which are expected to correspond to brain activity, rejecting clear artifacts based on the Fourier spectra and topographic distributions of the sources.

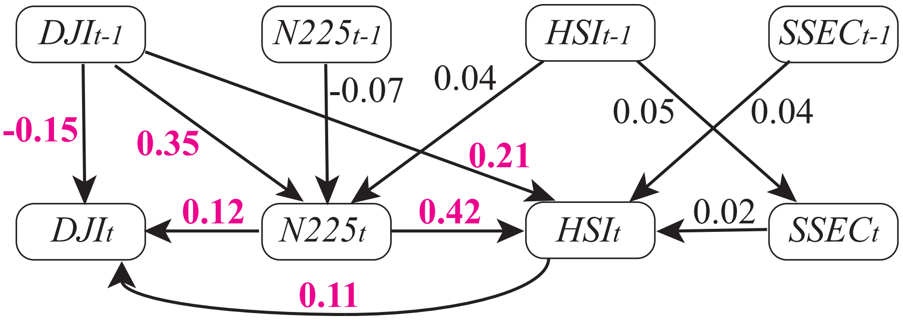


Figure 2: Results of application of our model to daily returns of the stock indices DJI, N225, HSI, and SSEC, with $k = 1$ lag. Large coefficients (greater than 0.1) are shown in bold and red.

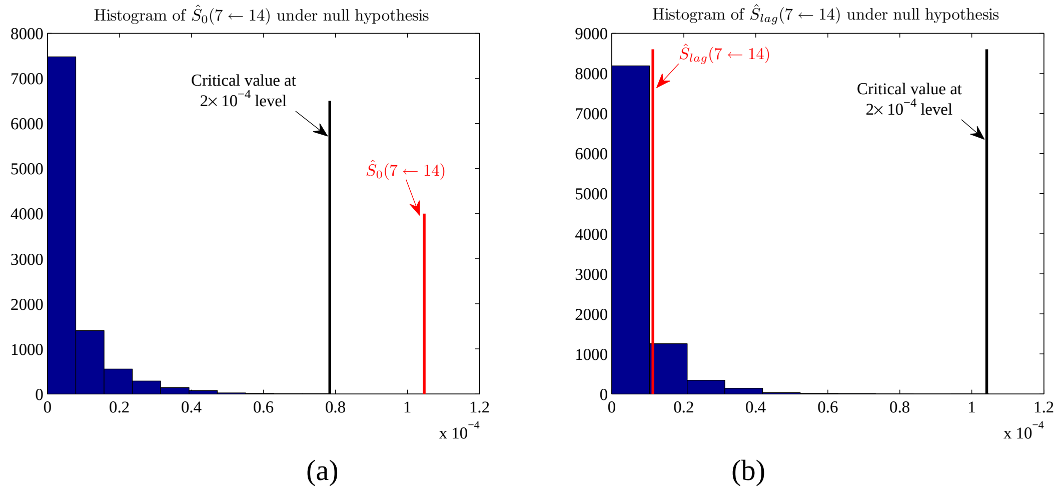
Next, we fitted an ordinary vector autoregressive model with 10 lags on the estimated sources, finding the corresponding innovation series which we denote by $y_i(t), i = 1, \ldots, 17$. Our goal was to analyze if there are some influences between the magnitudes of these innovations. We prefer to analyze the innovations because the innovations are approximately white both temporally and spatially, and thus we can analyze the magnitudes with no contamination by linear (auto)correlations of the source signals. The autoregressive model order 10 was chosen because it was the smallest order that gave approximately white innovations.

We then fitted the SVAR model on the logarithmically transformed magnitudes $x_i(t) = \log(0.2 + |y_i(t)|), i = 1, \ldots, 17$. We determined the order k of our SVAR model by minimizing the AIC criterion (Akaike, 1973), which is the negative log-likelihood of the MBD model plus a term measuring the complexity of the model. The log-likelihood involves the densities of the MBD outputs $\hat{e}_i(t)$, which were modelled by a mixture of three Gaussians. From the candidate orders between 0 and 20, we found that $k = 2$ gave the minimum AIC.

After finding the estimate of the coefficients $\hat{\mathbf{B}}_\tau, \tau = 0, 1, 2$ with the MBD-based approach, one can easily calculate the estimates of the statistics $S_0(i \leftarrow j)$ and $S_{lag}(i \leftarrow j)$. The bootstrapping approach given in Section 6 was used to evaluate if these estimated statistics are significant. Here we need to test multiple hypotheses simultaneously; to reduce the type I error, we adopted the Bonferroni correction (Shaffer, 1995) for multiple testing correction. We used the significance level 5%. For both the instantaneous and lagged effects, one needs to perform $17 \times 16 = 272$ tests; therefore, the significance level for each individual test is then $0.05/272 \approx 2 \times 10^{-4}$. We used 10^4 replications for the bootstrapping.

For illustration, we give the empirical distribution of the statistics $S_0(7 \leftarrow 14)$ and $S_{lag}(7 \leftarrow 14)$, as well as their estimated values for the original series $x_i(t)$, in Fig. 3. Clearly $\hat{S}_0(7 \leftarrow 14)$ is significant, while $\hat{S}_{lag}(7 \leftarrow 14)$ is not.

Fig. 4 shows the resulting diagram of causal analysis with instantaneous effects between the magnitudes of the selected MEG sources, with the influences significant at 5% level (corrected for multiple testing). What we see is that the connections tend to be strong between sources which are close to each other. For example, the occipitoparietal sources such as #1, #2, #3, #8, and #11 have strong interconnections. Some perirolandic sources such as #5, #7, #10, and #14 are also interconnected. Sources #4 and #16 seems to mediate between these two groups.



(a) (b)

Figure 3: Illustration of the empirical distribution of the statistics under the null hypothesis obtained by bootstrapping. (a) For the statistic $S_0(7 \leftarrow 14)$. (b) For $S_{lag}(7 \leftarrow 14)$.

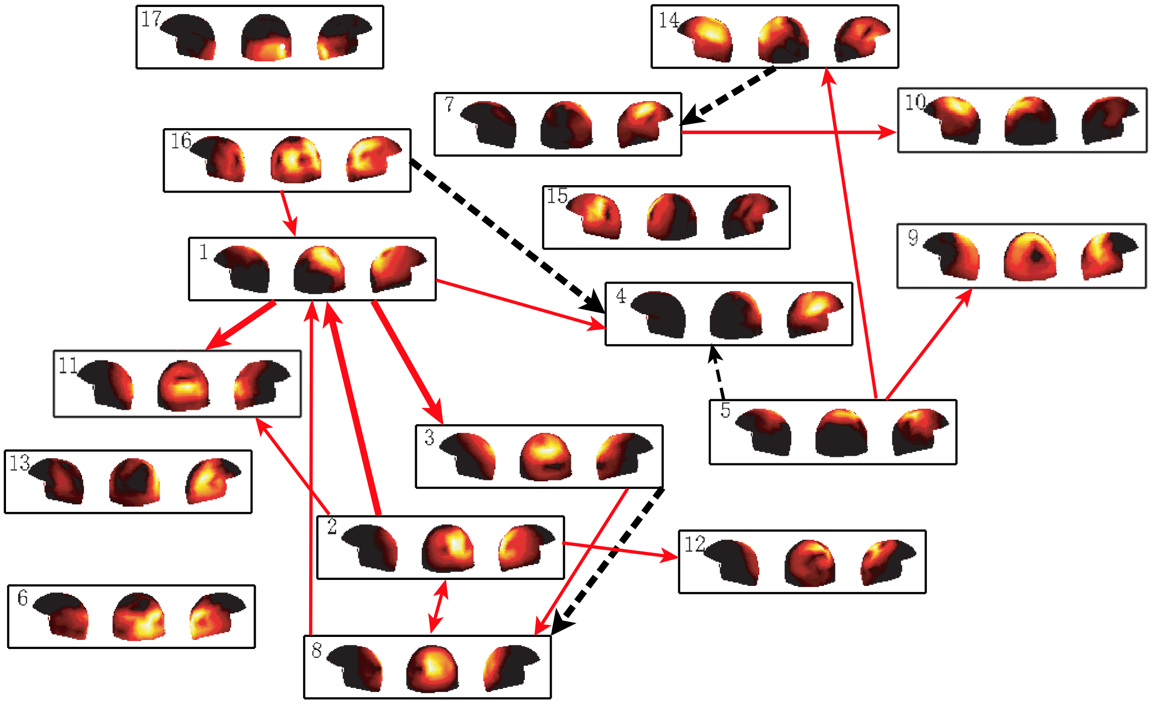


Figure 4: Results of application of our model on the log-magnitudes of the MEG sources (significant at 5% level, corrected for multiple testing). Black dashed line: instantaneous effect. Red solid line: lagged effect. The thickness of the lines indicates the strength of the influences.

9. Extensions of the Model

We have here assumed that $\mathbf{B}_0$ is acyclic, as is typical in causal analysis. However, this assumption is only made because we do not know very well how to estimate a linear non-Gaussian Bayesian network (or SEM) in the cyclic case. If we have a method which can estimate cyclic models, we do not need the assumption of acyclicity in our combined model either; see Lacerda et al. (2008) for one proposal. We could just use such a new method in our two-stage method instead of LiNGAM, and nothing else would be changed. However, development of methods for estimating cyclic models is orthogonal to the main contribution of our paper in the sense that we can use any such new method to estimate the instantaneous model in our framework.

In formulating the likelihood we had to assume that the $\mathbf{e}(t)$ are independent and identically distributed for different time points. However, in our two-stage estimation method, no such assumption was needed to guarantee consistency. In particular, the $\mathbf{e}(t)$ can be heteroscedastic, as long as $\mathbf{e}(t)$ and $\mathbf{e}(t')$ are uncorrelated for $t \neq t'$. In such a case, it might also be advantageous to change the LiNGAM estimation method so that the ICA part is replaced by methods estimating (4) explicitly based on temporal heteroscedasticity (Matsuoka et al., 1995; Hyvärinen, 2001; Pham and Cardoso, 2001); this is quite straightforward and necessitates no further changes in the method.

An interesting class of methods which is related to ours has been recently proposed by Gómez-Herrero et al. (2008). The idea is to combine blind source separation with a linear autoregressive model of the latent sources. The estimation of such a model can be accomplished by methods which are quite similar to our estimation methods, see also Haufe et al. (2009). However, the interpretation of the model is very different since, first, Gómez-Herrero et al. (2008) separate linear sources and analyze their (causal) connections whereas we analyze connections between the observed variables, and second, we estimate instantaneous causal influences whereas Gómez-Herrero et al. (2008) only estimate lagged ones.

10. Conclusion

We showed how non-Gaussianity enables estimation of a causal discovery model in which the linear effects can be either instantaneous or time-lagged. Like in the purely instantaneous case (Shimizu et al., 2006), non-Gaussianity makes the model identifiable without explicit prior assumptions on existence or non-existence of given causal effects. The theoretical developments are closely related to independent component analysis. The classic assumption of acyclicity was made, although it may not be necessary. From the practical viewpoint, an important implication is that considering instantaneous effects changes the coefficient of the time-lagged effects as well. We proposed methods for computationally efficient estimation of the model, as well as for sparsification and testing of the model coefficients.

Acknowledgments

We are grateful to Pavan Ramkumar, Lauri Parkkonen, and Riitta Hari for the MEG data. This work was partly funded by the Academy of Finland Centre-of-Excellence in Algorithmic Data Analysis.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proc. 2nd International Symposium on Information Theory*, pages 267–281, 1973.

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

E. M. L. Beale and C. L. Mallows. Scale mixing of symmetric distributions with zero means. *The Annals of Mathematical Statistics*, 30(4):1145–1151, 1959.

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

A. Chen and P. J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34 (6):2824–2855, 2006.

A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms*. Wiley, 2002.

P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

S. Demiralp and K. D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65 (supplement):745–767, 2003.

C. Diks and J. DeGoede. A general nonparametric bootstrap test for granger causality. In H. Broer and B. Krauskopf nd G. Vegter, editors, *Global Analysis of Dynamical Systems*, pages 391–403 (Chapter 16). Taylor & Francis, London, 2001.

R. F. Engle, editor. *ARCH: Selected Readings*. Oxford University Press, 1995.

K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J.L. Cantero. Measuring directional coupling between EEG sources. *NeuroImage*, 43:497–508, 2008.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

R. Hari and R. Salmelin. Human cortical oscillations: a neuromagnetic view through the skull. *Trends in Neuroscience*, 20:44–49, 1997.

S. Haufe, R. Tomioka, G. Nolte, K.-R. Müller, and M. Kawanabe. Modeling sparse connectivity between underlying brain sources for EEG/MEG. 2009. Arxiv preprint.

P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, 2001.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

A. Hyvärinen, S. Shimizu, and P. O. Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity. In *Proc. Int. Conf. on Machine Learning (ICML2008)*, pages 424–431, Helsinki, Finland, 2008.

A. Hyvärinen, P. Ramkumar, L. Parkkonen, and R. Hari. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. *NeuroImage*, 49(1):257–271, 2010.

J. Karvanen and V. Koivunen. Blind separation methods based on pearson system and its extensions. *Signal Processing*, 82(4):663–573, 2002.

J. Kim, W. Zhu, L. Chang, P. M. Bentler, and T. Ernst. Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Human Brain Mapping*, 28:85–93, 2007.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Proc. 24th Conf. on Uncertainty in Artificial Intelligence (UAI2008)*, Helsinki, Finland, 2008.

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.

K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

A. Moneta and P. Spirtes. Graphical models for the identification of causal structures in multivariate time series models. In *Proc. Joint Conference on Information Sciences*, Kaohsiung, Taiwan, 2006.

R. Opgen-Rhein and K. Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(37), 2007.

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

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49(9):1837–1848, 2001.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725, 1997.

A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, 25(1):230–242, 2005.

J. P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

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

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

N. R. Swanson and C. W. J. Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregression. *J. of the Americal Statistical Association*, 92: 357–367, 1997.

J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77–94, 1992.

K. Zhang and A. Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Proc. European Conference on Machine Learning (ECML2009)*, pages 570–585, 2009.

K. Zhang, H. Peng, L. Chan, and A. Hyvärinen. ICA with sparse connections: Revisited. In *Proc. Int. Conference on Independent Component Analysis and Blind Signal Separation (ICA2009)*, pages 195–202, Paraty, Brazil, 2009.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1417–1429, 2006.