Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity

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

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

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

Abstract

Causal analysis of continuous-valued variables typically uses either autoregressive models or linear Gaussian Bayesian networks with instantaneous effects. Estimation of Gaussian Bayesian networks 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. We show that such a non-Gaussian model is identifiable without prior knowledge of network structure, and we propose an estimation method shown to be consistent. This approach also points out how neglecting instantaneous effects can lead to completely wrong estimates of the autoregressive coefficients.

1. Introduction

Analysis of causal influences or effects has become an important topic in machine learning (Pearl, 2000; Spirtes et al., 1993), and has numerous applications in, for example, neuroinformatics (Roebroeck et al., 2005; Kim et al., 2007) and bioinformatics (Opgen-Rhein & Strimmer, 2007). For continuous-valued variables, such an analysis can basically be 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 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 causal influences are instantaneous, leading to Bayesian networks or structural equation models (Bollen, 1989).

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,[1] 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 Bayesian network 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 (other than the ubiquitous assumption of a directed acyclic graph (DAG)) is needed to estimate the model.

Here, we consider the general case where causal influences can occur either instantaneously or with considerable time lags. Such a model is called the structural vector autoregressive (SVAR) model in econometric theory, in which numerous attempts have been made for its estimation, see e.g. (Swanson & Granger, 1997; Demiralp & Hoover, 2003; Moneta & Spirtes, 2006). We propose to use non-Gaussianity to estimate the model. We show that this variant of the model is iden-

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

[1] Identifiability is here used in the classic statistical sense: a model is identifiable if no two different values of the parameter vector give the same distribution for the observed data.

tifiable without any other restrictions than acyclicity. 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 a computational method 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 leads to a generalization of the well-known Granger causality measure.

The paper is structured as follows. We first define the model and discuss its relation to other models in Section 2. In Section 3 we propose an estimation method, show its consistency, and discuss an intuitive interpretation of the method. Section 4 contains some theoretical examples and a theorem on how including instantaneous effects in the model changes the resulting interpretations. The resulting generalization of Granger causality is discussed in Section 5. The validity of the estimation method is demonstrated by simulations on artificial data in Section 6, and experiments on financial and neuroscientific data in Section 7. Section 8 concludes the paper.

2. Model combining lagged and instantaneous effects

2.1. Definition and assumptions

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 variables (time series) and t is the time index. All the variables are collected into a single vector $\mathbf{x}(t)$. Denote by k the number of time-delays used, i.e. the order of the autoregressive model. Denote by $\mathbf{B}_\tau$ the $n \times n$ matrix of the causal effects between the variables x_i with time lag $\tau, \tau = 0 \ldots k$.

The causal dynamics in our model are a combination of autoregressive and structural-equation models. The model is defined as

$$\mathbf{x}(t) = \sum_{\tau=0}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t) \qquad (1)$$

where the $e_i(t)$ are random processes modelling the external influences or "disturbances". We make the following assumptions on the external influences $e_i(t)$. First, they are mutually *independent*, and *temporally uncorrelated*, which are typical assumptions in autoregressive models. Second, they are assumed to be *non-Gaussian*, which is an important assumption which distinguishes our model from classic models, whether autoregressive models, structural-equation models, or Bayesian networks.

Further, we assume that the matrix modelling instantaneous effects, $\mathbf{B}_0$, corresponds to an *acyclic* graph, as is typical in causal analysis, but this may not be strictly necessary as will be discussed below. The acyclicity is equivalent to the existence of a permutation matrix $\mathbf{P}$, which corresponds to an ordering of the variables x_i, such that the matrix $\mathbf{P}\mathbf{B}_\mathbf{0}\mathbf{P}^T$ 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.

2.2. Relation to other models

This 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, i.e. $k = 0$, 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. In this sense the model is similar to independent component analysis, which solves the unidentifiability 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.

On the other hand, if the matrix $\mathbf{B}_0$ has all zero entries, the model in Equation (1) 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 Equation (1).

We emphasize that our model is different from classic autoregressive models two important ways: First, the external influences $e_i(t)$ are non-Gaussian. Second, the lag variable τ takes the value 0 as well, which brings instantaneous effects into the model in the form of the matrix $\mathbf{B}_0$. A coefficient $\mathbf{B}_0(i, j)$ models the instantaneous effect of $x_j(t)$ on $x_i(t)$ as in a linear Bayesian network, or a structural equation model.

2.3. Causality vs. prediction

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.

Our model is intended to be superior in causal modelling. Causality has an obvious intuitive interpretation, which is typically formalized as the ability to predict the effect of possible new *interventions* on the system (Pearl, 2000). Thus, our model should be better in predicting effects of interventions, which is different from conventional time series prediction.

3. Estimation of the model

3.1. Combining least-squares estimation and LiNGAM

We propose the following method for estimating our model defined in Section 2.1. The method combines classic least-squares estimation of an autoregressive (AR) model with LiNGAM estimation:

1. Estimate a classic autoregressive model for the data
$$\mathbf{x}(t) = \sum_{\tau=1}^{k} \mathbf{M}_\tau \mathbf{x}(t-\tau) + \mathbf{n}(t) \qquad (2)$$
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, i.e. estimates of innovations $\mathbf{n}(t)$
$$\hat{\mathbf{n}}(t) = \mathbf{x}(t) - \sum_{\tau=1}^{k} \hat{\mathbf{M}}_\tau \mathbf{x}(t-\tau) \qquad (3)$$

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
$$\hat{\mathbf{n}}(t) = \mathbf{B}_0 \hat{\mathbf{n}}(t) + \tilde{\mathbf{e}}(t) \qquad (4)$$

4. Finally, compute the estimates of the causal effect matrices $\mathbf{B}_\tau$ for $\tau > 0$ as
$$\hat{\mathbf{B}}_\tau = (\mathbf{I} - \hat{\mathbf{B}}_0)\hat{\mathbf{M}}_\tau \text{ for } \tau > 0 \qquad (5)$$

This estimation method is consistent,[2] as will be shown in Section 3.3. First, however, we show the derivation of Equation (5) and discuss its deep meaning.

3.2. Why autoregressive matrices change due to instantaneous influences

Equation (5) shows a remarkable 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, i.e. $\mathbf{B}_0 \neq 0$; otherwise, the estimates are not changed.

Why do we have (5)? This is because from (1) we have
$$(\mathbf{I} - \mathbf{B}_0)\mathbf{x}(t) = \sum_{\tau=1}^{k} \mathbf{B}_\tau \mathbf{x}(t-\tau) + \mathbf{e}(t) \qquad (6)$$
and thus
$$\mathbf{x}(t) = \sum_{\tau=1}^{k} (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_\tau \mathbf{x}(t-\tau) + (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{e}(t) \qquad (7)$$
Comparing this with (2), 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 (5) is justified.

While this phenomenon is, in principle, well-known in econometric literature (Swanson & Granger, 1997; Demiralp & Hoover, 2003; Moneta & Spirtes, 2006), Equation (5) 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 without strong prior assumptions on $\mathbf{B}_0$.

3.3. Consistency and identifiability

The consistency of our method relies on two facts. First, in the estimation of an AR model as in (2), it is not necessary that the innovation vector $\mathbf{n}(t)$ has independent or even uncorrelated elements (for fixed

[2] Consistency means classic statistical consistency, i.e. the estimator converges in probability to the right parameter values when the data follows the model and sample size grows infinite.

t); least-squares estimation will still be consistent, as is well known. Thus, least-squares estimation of (2), combined with (5), gives consistent estimators of $\mathbf{B}_\tau$ for $\tau \geq 1$, provided we have a consistent estimator of $\mathbf{B}_0$. Second, comparison of (7) with (2) shows that the residuals $\hat{\mathbf{n}}(t)$ are, asymptotically, of the form $(\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{e}(t)$. This means

$$\hat{\mathbf{n}}(t) = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{e}(t) \Leftrightarrow (\mathbf{I} - \mathbf{B}_0)\hat{\mathbf{n}}(t) = \mathbf{e}(t) \Leftrightarrow \hat{\mathbf{n}}(t) = \mathbf{B}_0\hat{\mathbf{n}}(t) + \mathbf{e}(t) \quad (8)$$

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.

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 in the cyclic case. Future work may produce methods which estimate cyclic models, and then we do not need the assumption of acyclicity in our combined model either. We could just use such a new method in Step 3 of the method instead of LiNGAM, and nothing else would be changed. Recent work in that direction is in (Lacerda et al., 2008); see also (Richardson & Spirtes, 1999) for older methods on Gaussian data.

3.4. Interpretation related to ICA of residuals

Another viewpoint on our model is analysis of the correlations of the innovations after estimating a classic AR model. Suppose we just estimate an AR model as in (2), and interpret the coefficients as causal effects. Such an interpretation more or less presupposes that the innovations n_i 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 present model. 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.

4. Interaction of instantaneous and lagged effects

Here we present some theoretical examples of how the instantaneous and lagged effects interact based on the formula in (5).

An instantaneous effect may seem to be lagged Consider first the case where the instantaneous and lagged matrices are as follows:

$$\mathbf{B}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix} \quad (9)$$

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

$$\mathbf{M}_1 = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1 = \begin{pmatrix} 0.9 & 0.9 \\ 0 & 0.9 \end{pmatrix} \quad (10)$$

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

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

$$\mathbf{B}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 0.9 \end{pmatrix} \quad (11)$$

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

$$\mathbf{M}_1 = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{B}_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0.9 & 0.9 & 0 \\ 0.9 & 0.9 & 0.9 \end{pmatrix} \quad (12)$$

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.

Causal ordering is not changed 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 $i(j), j = 1 \dots n$ of the variables such that no effect goes backward,*[3] *i.e.*

$$\mathbf{B}_\tau(i(j-\delta), i(j)) = 0 \text{ for } \delta > 0, \tau \geq 0, 1 \leq j \leq n \quad (13)$$

Then, the same property applies to the $\mathbf{M}_\tau, \tau \geq 1$ as well. Conversely, if there is an ordering such that (13) applies to $\mathbf{M}_\tau, \tau \geq 1$ and $\mathbf{B}_0$, then it applies to $\mathbf{B}_\tau, \tau \geq 1$ as well.

The proof of the theorem is based on the fact that 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 (5), which proves the first part of the theorem. The converse part follows from solving for $\mathbf{B}_\tau$ in (5) 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 (5) can change the causal interpretation of the results.

5. Towards a generalization of Granger causality

The classic interpretation of causality in instantaneous Bayesian network models would be that x_i causes x_j if the (j, i)-th coefficient in $\mathbf{B}_0$ is non-zero. In the time series context, this is related to Granger causality (Granger, 1969), which 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 propose a definition which combines the two aspects: 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.

A more general formulation of this definition, which is in line with the general formulation of Granger causality, is that the error in the "prediction" of $x_j(t)$ is reduced when $x_i(t-1), x_i(t-2), \dots$ *and* $x_i(t)$ are included in the set of predictors. Here, we use a rather unconventional definition of the word "prediction" because we include instantaneous effects.

6. Simulations

To verify the validity of our method, we first performed experiments with artificial data. In the experiments, we created data in the following manner using the LiNGAM code package[4]:

1. We randomly constructed a strictly lower-triangular matrix (i.e. zero entries above and on the diagonal), $\mathbf{B}_0$, for the instantaneous causal model so that the standard deviations of the innovations n_i owing to parent innovations will be in the interval $[0.5, 1.5]$. The number of observed time-series was $n = 10$. Both fully connected (no zeros in the strictly lower triangular part) and sparse networks (many zeros) were created. We also randomly selected the standard deviations of the external influences e_i from the interval $[0.5, 1.5]$.

2. Next, we generated data with various lengths of the time series (300, 500 and 1,000) by independently drawing the external influences e_i from various non-Gaussian distributions with zero mean and unit variance[5]. The values of the innovations n_i were generated according to the assumed instantaneous recursive process. This is straightforward because $\mathbf{B}_0$ is lower-triangular, so we just generate the n_i in the order $n_1, n_2 \dots$ as is typical in acyclic networks, e.g. (Shimizu et al., 2006).

3. We randomly permuted the order of the innovations n_i to hide the causal order with which the data was generated. We also permuted $\mathbf{B}_0$ as well

[3] In the purely instantaneous case, existence of such an ordering is equivalent to acyclicity of the effects as noted in Section 2.1.

[4] http://www.cs.helsinki.fi/group/neuroinf/lingam/

[5] We first generated a gaussian variable z with zero mean and unit variance and subsequently transformed it to a non-Gaussian variable by $e_i = \text{sign}(z)|z|^q$. The nonlinear exponent q was selected to lie in $[0.5, 0.8]$ or $[1.2, 2.0]$. The former gave a sub-gaussian variable, and the latter a super-gaussian variable. Finally, the transformed variable was standardized to have zero mean and unit variance.

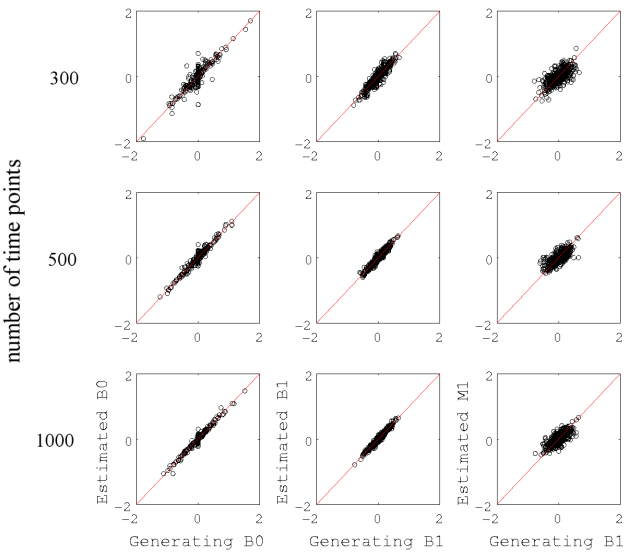

Figure 1. Simulations on artificial data. Left column: Scatterplots of the estimated elements of $\mathbf{B}_0$ versus the generating values. Center column: Scatterplots of the estimated elements of $\mathbf{B}_1$ versus the generating values. Right column: Scatterplots of the estimated elements of $\mathbf{M}_1$ versus those of $\mathbf{B}_1$.The number of observed signals was 10. Five data sets were generated for each scatterplot.

as the variances of the external influences e_i to match the new order.

4. We randomly generated a first-order autoregressive matrix $\mathbf{M}_1$ so that the spectral norm of the matrix was less than 0.99 to ensure the stability of the autoregressive process.

5. The values of the observed signals $x_i(t)$ were generated according to the assumed first-order autoregressive process.

6. Finally, we fed the data to our estimation method. Here we told the method that the generating autoregressive order was 1.

Figure 1 gives the scatterplots of the elements of the estimated parameters versus the generating ones. The left column is for the scatterplots of the estimated causal effects in $\mathbf{B}_0$ versus the generating values. The center column is for the scatterplots of the estimated causal effects in $\mathbf{B}_1$ versus the generating values. The right column is for the scatterplots of the estimated autoregressive coefficients in $\mathbf{M}_1$ versus the generating values of the causal effects in $\mathbf{B}_1$ (here, the estimation was invalid because instantaneous effects were ignored).

For the scatterplots in the left and center columns, the estimation worked well when the sample size grew, as evidenced by the grouping of the data points onto the main diagonal, although for the small sample size 300 the estimation was often inaccurate. On the other hand, the scatterplots in the right column confirmed that the causal effects were not correctly estimated by the ordinary autoregressive coefficients when instantaneous influences existed since the data points were not very close to the main diagonal.

7. Experiments on real data

7.1. Financial data

As a first illustration of the applicability of the method on real data, we analyzed a dataset from a time series repository on the Internet.[6] The data consisted of two observed signals, x_1: weekly closing price of Toyota stock and x_2: weekly closing rate of exchange of Japanese Yen to U.S. Dollar in 2007. The number of time points was 50. The maximum, minimum and mean of x_1 were 8,230, 5,870 and 7,102 (JPY). Those of x_2 were 123.86, 108.51 and 117.72 (JPY).

We analyzed the data using our method with autoregressive order of 1. The estimated first-order autoregressive matrix $\mathbf{M}_1$ and residual correlation matrix were as follows:

$$\mathbf{M}_1 = \begin{pmatrix} 0.95 & -4.22 \\ 0.0008 & 0.78 \end{pmatrix} \tag{14}$$

$$\mathrm{corr}(\mathbf{n}) = \begin{pmatrix} 1.00 & 0.66 \\ 0.66 & 1.00 \end{pmatrix}$$

The relatively strong correlation between the residuals implied that there would be some dependency that had not been modeled by the AR model. Thus, we fitted the instantaneous causal model to the residuals, as proposed above. The estimated instantaneous causal effect matrix $\mathbf{B}_0$ and resulting lagged causal effect matrix $\mathbf{B}_1$ were as follows:

$$\mathbf{B}_0 = \begin{pmatrix} 0 & 56.04 \\ 0.0027 & 0 \end{pmatrix} \tag{15}$$

$$\mathbf{B}_1 = \begin{pmatrix} 0.91 & -48.01 \\ -0.0018 & 0.79 \end{pmatrix} \tag{16}$$

The matrix $\mathbf{B}_0$ is very close to be upper-triangular, which implied that the model was really acyclic (because switching the order of the variables would make $\mathbf{B}_0$ lower-triangular). Further, the instantaneous effect $x_2 \to x_1$ in $\mathbf{B}_0$ was one order of magnitude larger than the lagged effect in $\mathbf{M}_1$ and thus the lagged co-

[6] Yahoo! Japan Finance: http://quote.yahoo.co.jp/

efficients in $\mathbf{M}_1$ are quite different from those in $\mathbf{B}_1$, due to the formula in (5).

Figure 2 shows a graphical representation of the estimated model for financial data. First, it implies that a higher value of the yen (x_2) had a negative lagged effect (-48.01) on the price of Toyota stock (x_1). This would be reasonable since Toyota sells many cars abroad, and a higher value of the yen would increase the cost price and decrease the earning. Interestingly, it was also implied that a higher value of the yen had a positive instantaneous effect (56.04) on the price of Toyota stock. In other words, for weeks where values of the yen one week before were the (approximately) same, if the yen got more expensive (due to some reason other than the value of the yen one week before, perhaps a U.S. recession, for example) then the price of Toyota stock would get more expensive. It would be interesting to further study the economic mechanism with more extensive data.

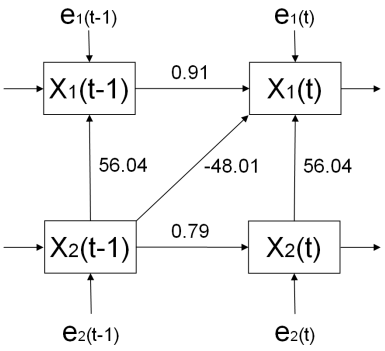

Figure 2. A graphical representation of the model estimated in Section 7.1. The x_1 and x_2 denote weekly closing price of Toyota stock in 2007 and weekly closing rate of exchange of Japanese Yen to U.S. Dollar in 2007, respectively. The arrow from $x_1(t-1)$ to $x_2(t)$ was omitted since the estimated strength was very close to zero (-0.0018).

7.2. Magnetoencephalographic data

As a second 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) in a magnetically shielded room at the Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology. The sampling frequency was 600 Hz. The measurements consisted of 300 seconds of resting state brain activity from the experiment of (Ramkumar et al., 2007). The subject was sitting with eyes closed, and did not perform any specific task nor was there any specific sensory stimulation. The channels were first linearly projected to the signal space to reduce noise (Uusitalo & Ilmoniemi, 1997). In this illustrative experiment, we only consider a single (gradiometer) channel in the right occipital cortex near the midline.

We considered the interaction of about 10 Hz (alpha) and about 20 Hz (beta) oscillations commonly observed in electromagnetic recordings of spontaneous brain activity. We first computed the amplitudes of the oscillations by dividing the data into windows of length of 0.25 seconds, performing fast Fourier transform inside each of them, and computing the total Fourier amplitudes (unweighted Euclidean norm of the Fourier coefficients) in the frequency ranges of $8 \ldots 12$Hz (alpha range, denoted by x_1) and $15 \ldots 25$Hz (beta range, denoted by x_2). Thus we obtained two time series of 1,200 points.

We fitted our model, with autoregressive order of 1 to the data. The obtained matrices are

$$\mathbf{M}_1 = \begin{pmatrix} 0.23381 & 0.14551 \\ 0.10838 & 0.14314 \end{pmatrix} \tag{17}$$

$$\mathbf{B}_0 = \begin{pmatrix} 0 & -0.65768 \\ 0.56722 & 0 \end{pmatrix} \tag{18}$$

$$\mathbf{B}_1 = \begin{pmatrix} 0.30509 & 0.23965 \\ -0.024244 & 0.060608 \end{pmatrix} \tag{19}$$

What we see is that the instantaneous model is far from trivial: the effects in $\mathbf{B}_0$ are relatively strong. This is also reflected in $\mathbf{B}_1$ which is now rather different from $\mathbf{M}_1$. Thus, the interpretation of the autoregressive matrices using just the autoregressive model (i.e. $\mathbf{M}_1$) or the combined model (i.e. $\mathbf{B}_1$) are quite different. In the classic autoregressive case (based on $\mathbf{M}_1$), the lagged effect $x_1 \to x_2$ is relatively strongly positive whereas in the combined model it is quite weak. In fact, that effect is now modelled as an instantaneous effect in $\mathbf{B}_0$. Even more interesting is that the instantaneous model has a strong negative effect $x_2 \to x_1$ which is not visible at all in the purely autoregressive matrix $\mathbf{M}_1$. Thus, the results illustrate how the interpretation of causal effects (and even of the lagged ones) can change drastically when including the instantaneous effects.

Using an autoregressive order of 2 did not change the results. We also ran the method many times to exclude the problem of the ICA estimation algorithm (used in LiNGAM estimation) getting stuck in local minima (Himberg et al., 2004), and the result was found to be robust with respect to that manipulation.

One problem with this experiment is that the causal model estimated by LiNGAM is far from acyclic. Here, we can justify the procedure by using the theory of cyclic model estimation proposed by (Lacerda et al., 2008); the estimation here gives the only "stable" model according to that theory. Performance of LiNGAM estimation methods in the case of cyclic models, and the possible need for new methods for estimating cyclic models are future research topics of great practical importance. However, as discussed above, they are separate from 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.

8. 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 classic assumption of acyclicity is sufficient although probably not necessary. From the practical viewpoint, an important implication is that considering instantaneous effects changes the coefficient of the time-lagged effects as well.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

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

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

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

Himberg, J., Hyvärinen, A., & Esposito, F. (2004). Validating the independent components of neuroimaging time-series via clustering and visualization. *NeuroImage*, *22*, 1214–1222.

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

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

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

Moneta, A., & Spirtes, P. (2006). Graphical models for the identication of causal structures in multivariate time series models. *Proc. Joint Conference on Information Sciences*. Kaohsiung, Taiwan.

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Ramkumar, P., Parkkonen, L. T., He, B. J., Raichle, M. E., Hämäläinen, M. S., & Hari, R. (2007). Identification of stimulus-related and intrinsic networks by spatial independent component analysis of MEG signals. Abstract presented at the Society for Neuroscience Meeting, San Diego, California.

Richardson, T. S., & Spirtes, P. (1999). Automated discovery of linear feedback models. In C. Glymour and G. Cooper (Eds.), *Computation, causation and discovery*, 253–302. The MIT Press.

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

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag.

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

Uusitalo, M. A., & Ilmoniemi, R. J. (1997). Signal-space projection method. *Med. Biol. Eng.*, *32*, 35–42.