Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components

Philipp Geiger[a] PGEIGER@TUEBINGEN.MPG.DE
Kun Zhang[a,b] KZHANG@TUEBINGEN.MPG.DE
Mingming Gong[c] GONGMINGNJU@GMAIL.COM
Dominik Janzing[a] JANZING@TUEBINGEN.MPG.DE
Bernhard Schölkopf[a] BS@TUEBINGEN.MPG.DE

[a]Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany
[b]Information Sciences Institute, University of Southern California, USA
[c]Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia

Abstract

A widely applied approach to causal inference from a time series X, often referred to as "(linear) Granger causal analysis", is to simply regress present on past and interpret the regression matrix $\hat{B}$ causally. However, if there is an unmeasured time series Z that influences X, then this approach can lead to wrong causal conclusions, i.e., distinct from those one would draw if one had additional information such as Z. In this paper we take a different approach: We assume that X together with some hidden Z forms a first order vector autoregressive (VAR) process with transition matrix A, and argue why it is more valid to interpret A causally instead of $\hat{B}$. Then we examine under which conditions the most important parts of A are identifiable or almost identifiable from only X. Essentially, sufficient conditions are (1) non-Gaussian, independent noise or (2) no influence from X to Z. We present two estimation algorithms that are tailored towards conditions (1) and (2), respectively, and evaluate them on synthetic and real-world data. We discuss how to check the model using X.

1. Introduction

Inferring the causal structure of a stochastic dynamical system from a time series of measurements is an important problem in many fields such as economics (Lütkepohl, 2006) and neuroscience (Roebroeck et al., 2005; Besserve et al., 2010).

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

In the present paper, we approach this problem as follows: We assume that the measurements are a finite sample from a random process $X = (X_t)_{t\in\mathbb{Z}}$ which, together with another random process $Z = (Z_t)_{t\in\mathbb{Z}}$, forms a first order vector autoregressive (VAR) process. That is, $(X, Z)^\top$ obeys

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \begin{pmatrix} B & C \\ D & E \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + N_t,$$

for all $t \in \mathbb{Z}$, some matrices B, C, D, E, and some i.i.d. $N_i, i \in \mathbb{Z}$. So far this is a purely statistical model. Now we additionally assume that the variables in Z correspond to real properties of the underlying system that are in principle measurable and intervenable. Based on this we consider B, C, D, E to have a *causal meaning*. More precisely, we assume that B's entries express the direct causal influences between the respective variables in X. And more generally, we assume that for all variables in $(X, Z)^\top$ the matrices B, C, D, E capture the respective direct and indirect causal influences. Note that in this sense C is particularly interesting because it tells which components of X are jointly influenced by an unmeasured quantity, i.e., have a *hidden confounder*, and how strong the influence is.

This way causal inference on X is reduced to a *statistical* problem: examining to what extent, i.e., under which assumptions, B as well as C, D, E are identifiable from the distribution of the process X, and how they can be estimated from a sample of X. It is worth mentioning that this approach can be justified in two different ways, following either (Granger, 1969) or (Pearl, 2000; Spirtes et al., 2000). We will briefly elaborate on this later (Section 4.2).

The first and main contribution of this paper is on the theoretical side: we present several results that show under which conditions B and C are identifiable or almost (i.e. up to a small number of possibilities) identifiable from only the distribution of X. Generally we assume that Z has at

most as many components as X. Theorem 1 shows that if the noise terms are non-Gaussian and independent, and an additional genericity assumption holds true, then B is uniquely identifiable. Theorem 2 states that under the same assumption, those columns of C that have at least two non-zero entries are identifiable up to scaling and permutation indeterminacies (because scale and ordering of the components of Z are arbitrary). Theorem 3 shows that regardless of the noise distribution (i.e., also in the case of Gaussian noise), if there is no influence from X to Z and an additional genericity assumption holds, then B is identifiable from the covariance structure of X up to a small finite number of possibilities. In Propositions 1 and 2 we prove that the additional assumptions we just called generic do in fact only exclude a Lebesgue null set from the parameter space.

The second contribution is a first examination of how the above identifiability results can be translated into estimation algorithms on finite samples of X. We propose two algorithms. Algorithm 1, which is tailored towards the conditions of Theorems 1 and 2, estimates B and C by approximately maximizing the likelihood of a parametric VAR model with a mixture of Gaussians as noise distribution. Algorithm 2, which is tailored towards the conditions of Theorem 3, estimates the matrix B up to finitely many possibilities by solving a system of equations somewhat similar to the Yule-Walker equations (Lütkepohl, 2006). Furthermore, we briefly examine how the model assumptions that we make can to some extend be checked just based on the observed sample of X. We examine the behavior of the two proposed algorithms on synthetic and real-world data.

It should be mentioned that probably the most widely applied approach to causal inference from time series data so far (Lütkepohl, 2006), which we refer to as *practical Granger causal analysis* in this paper (often just called "(linear) Granger causality"), is to simply perform a linear regression of present on past on the observed sample of X and then interpret the regression matrix causally. While this method may yield reasonable results in certain cases, it obviously can go wrong in others (see Section 4.3 for details). We believe that the approach presented in this paper may in certain cases lead to more valid causal conclusions.

The remainder of this paper is organized as follows. In Section 2 we discuss related work. In Section 3 we introduce notation and definitions for time series. In Section 4 we state the statistical and causal model that we assume throughout the paper. In Section 5 we introduce the so-called generalized residual. Section 6 contains the three main results on identifiability (Theorems 1 to 3) as well as arguments for the genericity of certain assumptions we need to make (Propositions 1 and 2). In Section 7 we present the two estimation algorithms and discuss model checking. Section 8 contains experiments for Algorithms 1 and 2. We conclude with Section 9.

2. Related Work

We briefly discuss how the present work is related to previous papers in similar directions.

Inference of properties of processes with hidden components: The work (Jalali & Sanghavi, 2012) also assumes a VAR model with hidden components and tries to identify parts of the transition matrix. However their results are based on different assumptions: they assume a "local-global structure", i.e., connections between observed components are sparse and each latent series interacts with many observed components, to achieve identifiability. The authors of (Boyen et al., 1999) - similar to us - apply a method based on expectation maximization (EM) to infer properties of partially observed Markov processes. Unlike us, they consider finite-state Markov processes and do not provide a theoretical analysis of conditions for identifiability. The paper (Etesami et al., 2012) examines identifiability of partially observed processes that have a certain tree-structure, using so-called discrepancy measures.

Harnessing non-Gaussian noise for causal inference: The paper (Hyvaerinen et al., 2010) uses non-Gaussian noise to infer instantaneous effects. In (Hoyer et al., 2008), the authors use the theory underlying overcomplete independent component analysis (ICA) (Kagan et al., 1973, Theorem 10.3.1) to derive identifiability (up to finitely many possibilities) of linear models with hidden variables, which is somewhat similar to our Theorem 1. However, there are two major differences: First, they only consider models which consist of finitely many observables which are mixtures of finitely many noise variables. Therefore their results are not directly applicable to VAR models. Second, they show identifiability only up to a finite number of possibilities, while we (exploiting the autoregressive structure) prove unique identifiability.

Integrating several definitions of causation: The work (Eichler, 2012) provides an overview over various definitions of causation w.r.t. time series, somewhat similar to but more comprehensive than our brief discussion in Sections 4.2 and 4.3.

3. Time Series: Notation and Definitions

Here we introduce notation and definitions w.r.t. time series. We denote multivariate *time series*, i.e., families of random vectors over the index set $\mathbb{Z}$, by upper case letters such as X. As usual, X_t denotes the t-th member of X, and X_t^k denotes the k-th component of the random vector X_t. Slightly overloading terminology, we call the univariate time series $X^k = (X_t^k)_{t\in\mathbb{Z}}$ the *k-th component of X*. By P_X we denote the distribution of the random process X, i.e., the joint distribution of all X_t, $t \in \mathbb{Z}$.

Given a K_X-variate time series X and a K_Z-variate time

series Z, $(X, Z)^\top$ denotes the $(K_X + K_Z)$-variate series

$$\left((X_t^1, \ldots, X_t^{K_X}, Z_t^1, \ldots, Z_t^{K_Z})^\top\right)_{t\in\mathbb{Z}}.$$

A K-variate time series W is a *vector autoregressive process (of order* 1*), or VAR process for short, with VAR transition matrix* A *and noise covariance matrix* Σ, if it allows a *VAR representation*, i.e.,

$$W_t = AW_{t-1} + N_t, \tag{1}$$

the absolute value of all eigenvalues of A is less than[1] 1, and N is an i.i.d. noise time series such that $\mathrm{Cov}(N_0) = \Sigma$. We say W is a *diagonal-structural VAR process* if in the above definition the additional condition is met that $N_0^1, \ldots, N_0^K$ are jointly independent.[2]

4. Statistical and Causal Model Assumptions

In this section we introduce the statistical model that we consider throughout the paper and discuss based on which assumptions its parameters can be interpreted causally. Moreover, we give an example for how practical Granger causal analysis can go wrong.

4.1. Statistical Model

Let K_X be arbitrary but fixed. Let X be a K_X-variate time series. As stated in Section 1, X is the random process from which we assume we measured a sample. In particular, the random variables in X have a meaning in reality (e.g., X_3^1 is the temperature measured in room 1 at time 3) and we are interested in the causal relations between these variables. Let X be related to a K-variate VAR process W, with transition matrix A, noise time series N, and noise covariance matrix Σ, and a K_Z-variate time series Z, as follows: $W = (X, Z)^\top$ and $K_Z \leq K_X$. Furthermore, let

$$A =: \begin{pmatrix} B & C \\ D & E \end{pmatrix}, \tag{2}$$

with B a $K_X \times K_X$ matrix. We call B, the most interesting part of A, the *structural matrix underlying* X. Furthermore, in case $C \neq 0$, we call Z a *hidden confounder*.

4.2. Causal Assumptions

As already mentioned in Section 1, throughout this paper we assume that there is an underlying system such that all variables in W correspond to actual properties of that system which are in principle measurable and intervenable. While we assume that a finite part of X was in fact measured (Section 4.1), Z is completely unmeasured. Furthermore we assume that the entries of A, in particular the submatrix B, capture the actual non-instantaneous causal influences between the variables in W. We also mentioned that there are two lines of thought that justify this assumption. We briefly elaborate on this here.

On the one hand, (Granger, 1969) proposed a definition of causation between observables which we will refer to as *Granger's ideal definition*. Assume the statistical model for the observed sample of X specified in Section 4.1. If we additionally assume that Z correctly models the whole rest of the universe or the "relevant" subpart of it, then according to Granger's ideal definition the non-instantaneous (direct) causal influences between the components of X are precisely given by the entries of B. But this implies that everything about B that we can infer from X can be interpreted causally, if one accepts Granger's ideal definition and the additional assumptions that are necessary (such as $K_Z \leq K_X$, which in fact may be a quite strong assumption of course). This is one way to justify our approach.

On the other hand, (Pearl, 2000) does not define causation based on measurables alone but instead formalizes causation by so-called structural equation models (SEMs) and links them to observable distributions via additional assumptions. In this sense, let us assume that W forms a causally sufficient set of variables, whose correct structural equations are given by the VAR equations (1), i.e., these equations represent actual causal influences from the r.h.s. to the l.h.s.[3] In particular these equations induce the correct (temporal) causal directed acyclic graph (DAG) for $(X, Z)^\top$. Then, essentially following the above mentioned author, everything about B that we can infer from the distribution of X can be interpreted causally. This is the other way to justify our approach (in case the requirement $K_Z \leq K_X$ and the other assumptions are met). It is important to mention that the usual interpretation of SEMs is that they model the mechanisms which generate the data and that they predict the outcomes of randomized experiments w.r.t. the variables contained in the equations.

4.3. Relation to Practical Granger Causal Analysis and How It Can Go Wrong

The above ideal definition of causation by Granger (Section 4.2) needs to be contrasted with what we introduced as "practical Granger causal analysis" in Section 1. In practical Granger causal analysis, one just performs a linear regression of present on past on the observed X and then interprets the regression matrix causally.[4] While making the

[1] We require all VAR processes to be stable (Lütkepohl, 2006).

[2] Note that the notion "diagonal-structural" is a special case of the more general notion of "structural" in e.g., (Lütkepohl, 2006).

[3] Note that here we ignore the fact that Pearl generally only considers models with finitely many variables while the process W is a family of infinitely many (real-valued) variables.

[4] We are aware that nonlinear models (Chu & Glymour, 2008) and nonparametric estimators (Schreiber, 2000) have been used to find temporal causal relations. In this paper we focus on the

ideal definition practically feasible, this may lead to wrong causal conclusions in the sense that it does not comply with the causal structure that we would infer given we had more information.[5]

Let us give an example for this. Let X be bivariate and Z be univariate. Moreover, assume

$$A = \left(\begin{array}{cc|c} 0.9 & 0 & 0.5 \\ 0.1 & 0.1 & 0.8 \\ \hline 0 & 0 & 0.9 \end{array}\right),$$

and let the covariance matrix of N_t be the identity matrix. To perform practical Granger causal analysis, we proceed as usual: we fit a VAR model on *only* X, in particular compute, w.l.o.g. assuming zero mean, the transition matrix by

$$B_{\text{pG}} := \mathbb{E}(X_t X_{t-1}^\top)\mathbb{E}(X_t X_t^\top)^{-1} = \begin{pmatrix} 0.89 & 0.35 \\ 0.08 & 0.65 \end{pmatrix} \tag{3}$$

(up to rounding) and interpret the coefficients of B_{pG} as causal influences. Although, based on A, X_t^2 does in fact not cause X_{t+1}^1, B_{pG} suggests that there is a strong causal effect $X_t^2 \to X_{t+1}^1$ with the strength 0.35. It is even stronger than the relation $X_t^1 \to X_{t+1}^2$, which actually exists in the complete model with the strength 0.1.

5. The Generalized Residual: Definition and Properties

In this section we define the generalized residual and discuss some of its properties. The generalized residual is used in the proofs of the three main results of this paper, Theorems 1 to 3.

For any $K_X \times K_X$ matrices U_1, U_2 let

$$R_t(U_1, U_2) := X_t - U_1 X_{t-1} - U_2 X_{t-2}.$$

We call this family of random vectors *generalized residual*. Furthermore let

$$M_1 := \mathbb{E}\left[W_t \cdot (X_t^\top, X_{t-1}^\top)\right].$$

In what follows, we list some simple properties of the generalized residual. Proofs can be found in (Geiger et al., 2015, Section A).

Lemma 1. *We have*

$$\begin{aligned} R_t(U_1, U_2) = {} & (B^2 + CD - U_1 B - U_2) X_{t-2} \\ & + (BC + CE - U_1 C) Z_{t-2} \\ & + (B - U_1) N_{t-1}^X + C N_{t-1}^Z + N_t^X, \end{aligned} \tag{4}$$

if $K > K_X$. In case $K = K_X$, the same equation holds except that one sets $C := D := E := 0$.

Lemma 2. *If (U_1, U_2) satisfies the equation*

$$(U_1, U_2)\begin{pmatrix} B & C \\ \mathrm{I} & 0 \end{pmatrix} = \left(B^2 + CD, BC + CE\right), \tag{5}$$

then $R_t(U_1, U_2)$ is independent of $(X_{t-2-j})_{j=0}^\infty$, and in particular, for $j \geq 0$,

$$\mathrm{Cov}(R_t(U_1, U_2), X_{t-2-j}) = 0. \tag{6}$$

Let $\Gamma_i^X := \mathrm{Cov}(X_t, X_{t-i})$ for all i. That is, Γ_i^X are the *autocovariance matrices* of X. Note that equation (6), for $j = 0, 1$, can equivalently be written as the single equation

$$(U_1, U_2)\begin{pmatrix} \Gamma_1^X & \Gamma_2^X \\ \Gamma_0^X & \Gamma_1^X \end{pmatrix} = \left(\Gamma_2^X, \Gamma_3^X\right). \tag{7}$$

Keep in mind that, as usual, we say a $m \times n$ matrix has *full rank* if its (row and column) rank equals $\min\{m, n\}$.

Lemma 3. *Let M_1 have full rank. If (U_1, U_2) satisfies equation (6) for $j = 0, 1$, then it satisfies equation (5).*

Lemma 4. *If $K = K_X$ or if C has full rank, then there exists (U_1, U_2) that satisfies equation (5).*

6. Theorems on Identifiability and Almost Identifiability

This section contains the main results of the present paper. We present three theorems on identifiability and almost identifiability of B and C (defined in Section 4.1), respectively, given X and briefly argue why certain assumptions we have to make can be considered as generic. Recall the definition of the matrix M_1 in Section 5. Note that the following results show (almost) identifiability of B for all numbers K_Z of hidden components *simultaneously*, as long as $0 \leq K_Z \leq K_X$ (which contains the case of no hidden components as a special case).

6.1. Assuming Non-Gaussian, Independent Noise

We will need the following assumptions for the theorems.

Assumptions. *We define the following abbreviations for the respective subsequent assumptions.*

A1: *All noise terms N_t^k, $k = 1, \ldots, K, t \in \mathbb{Z}$, are non-Gaussian.*

A2: *W is a diagonal-structural VAR process (as defined in Section 3).*

G1: *C (if it is defined, i.e., if $K > K_X$) and M_1 have full rank.*

linear case.

[5] Obviously, if one is willing to assume that X is causally sufficient already, then the practical Granger causation can be justified along the lines of Section 4.2.

(We will discuss the genericity of G1 in Section 6.3.)

The following definition of F_1 is not necessary for an intuitive understanding, but is needed for a precise formulation of the subsequent identifiability statements. Let F_1 denote the set of all K'-variate VAR processes W' with $K_X \leq K' \leq 2K_X$ (i.e. W has at most as many hidden components as observed ones), which satisfy the following properties w.r.t. N', C', M_1' (defined similarly to N, C, M_1 in Section 4): assumptions A1, A2 and G1 applied to N', C', M_1' (instead of N, C, M_1) hold true.

Theorem 1. *If assumptions A1, A2 and G1 hold true, then B is uniquely identifiable from only P_X.*

That is: There is a map f such that for each $W' \in F_1$, and X' defined as the first K_X components of W', $f(P_{X'}) = B'$ iff B' is the structural matrix underlying X'.

A detailed proof can be found in (Geiger et al., 2015, Section B.1). The idea is to chose U_1, U_2 such that $R_t(U_1, U_2)$ is a linear mixture of only *finitely* many noise terms, which is possible based on Lemmas 1 to 4. Then, using the identifiability result underlying overcomplete ICA (Kagan et al., 1973, Theorem 10.3.1), the structure of the mixing matrix of $(R_t(U_1, U_2), R_{t-1}(U_1, U_2))^\top$ allows to uniquely determine B from it.

Again using (Kagan et al., 1973, Theorem 10.3.1), one can also show the following result. For a matrix M let $S(M)$ denote the set of those columns of M that have at least two non-zero entries, and if M is not defined, let $S(M)$ denote the empty set. A proof can be found in (Geiger et al., 2015, Section B.2).

Theorem 2. *If assumptions A1, A2 and G1 hold true, then the set of columns of C with at least two non-zero entries is identifiable from only P_X up to scaling of those columns.*

In other words: There is a map f such that for each $W' \in F_1$ with K' components, X' defined as the first K_X components of W', and C' defined as the upper right $K_X \times (K' - K_X)$ submatrix of the transition matrix of W', $f(P_{X'})$ coincides with $S(C')$ up to scaling of its elements.

6.2. Assuming $D = 0$

In this section we present a theorem on the almost identifiability of B under different assumptions. In particular, we drop the non-Gaussianity assumption. Instead, we make the assumption that Z is not influenced by X, i.e., $D = 0$.

Given $U = (U_1, U_2)$, let

$$T_U(Q) := Q^2 - U_1 Q - U_2, \tag{8}$$

for all square matrices Q that have the same dimension as U_1. Slightly overloading notation, we let $T_U(\alpha) := T_U(\alpha \mathrm{I})$ for all scalars α. Note that $\det(T_U(\alpha))$ is a univariate polynomial in α.

We will need the following assumptions for the theorem.

Assumptions. *We define the following abbreviations for the respective subsequent assumptions.*

A3*:* $D = 0$.

G2*: The transition matrix A is such that there exists $U = (U_1, U_2)$ such that equation (5) is satisfied and $\det(T_U(\alpha))$ has $2K_X$ distinct roots.*

(We will discuss the genericity of G2 in Section 6.3.)

The following definition of F_2 is not necessary for an intuitive understanding, but is needed for a precise formulation of the subsequent identifiability statement. Let F_2 denote the set of all K'-variate VAR processes W' with $K_X \leq K' \leq 2K_X$, which satisfy the following properties w.r.t. N', A', C', D', M_1' (defined similarly to N, A, C, D, M_1 in Section 4): assumptions A3, G1 and G2 applied to N', A', C', D', M_1' (instead of N, A, C, D, M_1) hold true.

Theorem 3. *If assumptions A3, G1 and G2 hold true, then B is identifiable from only the covariance structure of X up to $\binom{2K_X}{K_X}$ possibilities.*

In other words: There is a map f such that for each $W' \in F_2$, and X' defined as the first K_X components of W', $f(X')$ is a set of at most $\binom{2K_X}{K_X}$ many matrices, and $B' \in f(P_{X'})$ for B' the structural matrix underlying X'.

A detailed proof can be found in (Geiger et al., 2015, Section B.3). The proof idea is the following: Let L denote the set of all $(U, \tilde{B})$, with $U = (U_1, U_2)$, that satisfy equation (6) for $j = 0, 1$, as well as the equation

$$T_U(\tilde{B}) = 0, \tag{9}$$

and meet the condition that $\det(T_U(\alpha))$ has $2K_X$ distinct roots. L is non-empty and (U, B) is an element of it, for the true B and some U, due to Lemmas 2 to 4. But L is only defined based on the covariance of X and has at most $\binom{2K_X}{K_X}$ elements (based on (J. E. Dennis et al., 1976)).

Note the similarity between equation (6), or its equivalent, equation (7), and the well-known Yule-Walker equation (Lütkepohl, 2006). The Yule-Walker equation (which is implicitly used in equation (3)) determines B uniquely under some genericity assumption and given $C = 0$.

6.3. Discussion on the Genericity of Assumptions G1 and G2

In this section we want to briefly argue why the assumptions G1 and G2 are generic. A detailed elaboration with precise definitions and proofs can be found in (Geiger et al., 2015, Section C). The idea is to define a natural parametrization of (A, Σ) and to show that the restrictions that assumptions G1 and G2, respectively, impose on

(A, Σ) just exclude a Lebesgue null set in the natural parameter space and thus can be considered as generic.

In this section, let K such that $K_X \leq K \leq 2K_X$ be arbitrary but fixed. Let λ_k denote the k-dimensional Lebesgue measure on $\mathbb{R}^k$.

Let Θ_1 denote the set of all possible parameters (A', Σ') for a K-variate VAR processes W' that additionally satisfy assumption A2, i.e., correspond to structural W'. Let S_1 denote the subset of those $(A', \Sigma') \in \Theta_1$ for which also assumption G1 is satisfied. And let g denote the natural parametrization of Θ_1 which is defined in (Geiger et al., 2015, Section C.1).

Proposition 1. *We have* $\lambda_{K^2+K}\left(g^{-1}(\Theta_1 \setminus S_1)\right) = 0$.

A proof can be found in (Geiger et al., 2015, Section C.1). The proof idea is that $g^{-1}(\Theta_1 \setminus S_1)$ is essentially contained in the union of the root sets of finitely many multivariate polynomials and hence is a Lebesgue null set.

Let Θ_2 denote the set of all possible parameters (A', Σ') for the K-variate VAR processes W that additionally satisfy assumption A3, i.e., are such that the submatrix D of A is zero. Let S_2 denote the subset of those $(A', \Sigma') \in \Theta_2$ for which also assumptions G1 and G2 are satisfied. Let h denote the natural parametrization of Θ_1 which is defined in (Geiger et al., 2015, Section C.2). A proof for the following proposition (which is based on a similar idea as that of Proposition 1) can also be found in (Geiger et al., 2015, Section C.2).

Proposition 2. *We have* $\lambda_{2K^2-K_XK_Z}\left(h^{-1}(\Theta_2 \setminus S_2)\right) = 0$.

7. Estimation Algorithms

In this section we examine how the identifiability results in Section 6 can be translated into estimators on finite data. We propose two algorithms.

7.1. Algorithm Based on Variational EM

Here we present an algorithm for estimating B and C which is closely related to Theorems 1 and 2. Keep in mind that the latter theorem in fact only states identifiability for $S(C)$ (defined in Section 6.2), up to scaling, not for the exact C. The idea is the following: We transform the model of X underlying these theorems (i.e. the general model from Section 4.1 together with assumptions A1, A2 and G1 from Section 6.1) into a parametric model by assuming the noise terms N_t^k to be mixtures of Gaussians.[6] Then we estimate all parameters, including B and C, by approximately maximizing the likelihood of the given sample of X using a variational expectation maximization (EM) approach similar to the one in (Oh et al., 2005). (Directly maximizing the likelihood is intractable due to the hidden variables (Z and mixture components) that have to be marginalized out.) Let $y_{1:L}$ be shorthand for $(y_1, \ldots, y_L)$. The estimator is outlined by Algorithm 1, where (V_t^X, V_t^Z) with values (v_t^X, v_t^Z) denote the vectors of mixture components for N_t^X and N_t^Z, respectively; $q^j(z_{1:L}, v_{1:L}^X, v_{1:L}^Z | x_{1:L})$ the true posterior of $Z_{1:L}, V_{1:L}^X, V_{1:L}^Z$ under the respective parameter vector θ^j (which comprises A, Σ as well as the Gaussian mixture parameters) at step j; and $\mathcal{L}$ the variational lower bound. The detailed algorithm can be found in (Geiger et al., 2015, Section D). Note that, if needed, one may use cross validation as a heuristic to determine K_Z and the number of Gaussian mixture components.

[6] Obviously, Theorems 1 and 2 also imply identifiability of B and (up to scaling) $S(C)$ for this parametric model. We conjecture that this implies consistency of the (non-approximate) maximum likelihood estimator for that model under appropriate assumptions.

Algorithm 1 Estimate B, C using variational EM

1: **Input:** Sample $x_{1:L}$ of $X_{1:L}$.
2: Initialize the transition matrix and the parameters of the Gaussian mixture model, denoted as θ^0, set $j \leftarrow 0$.
3: **repeat**
4: **E step:** Evaluate

$$q^j(z_{1:L}, v_{1:L}^X, v_{1:L}^Z) = q^j(z_{1:L}) q^j(v_{1:L}^X) q^j(v_{1:L}^Z),$$

which is the variational approx. to the true posterior $q^j(z_{1:L}, v_{1:L}^X, v_{1:L}^Z | x_{1:L})$, by maximizing the variational lower bound, i.e., $q^j = \arg\max_q \mathcal{L}(q, \theta^j)$.
5: **M step:** Evaluate $\theta^{j+1} = \arg\max_\theta \mathcal{L}(q^j, \theta)$.
6: $j \leftarrow j + 1$.
7: **until** convergence
8: **Output:** The final θ^j, containing the estimated B, C.

Algorithm 2 Estimate B using covariance structure

1: **Input:** Sample $x_{1:L}$ of $X_{1:L}$.
2: Solve the equation (7), with Γ_i^X replaced by $\hat{\Gamma}_i^X$. Let $(\hat{U}_1, \hat{U}_2)$ denote the solution.
3: Solve equation (9) with $U := (\hat{U}_1, \hat{U}_2)$ for $\tilde{B}$. Let $\hat{B}_1, \ldots, \hat{B}_n$ denote the solvents.
4: **Output:** $\hat{B}_1, \ldots, \hat{B}_n$.

7.2. Algorithm Based on the Covariance Structure

Now we present an algorithm, closely related to Theorem 3, for estimating B up to finitely many possibilities. It relies on the proof idea of that theorem, as we outlined it at the end of Section 6.2, and it is meant to be applied for cases where the conditions of that theorem are met. It uses only the estimated autocovariance structure of X. Keep in mind that $\hat{\Gamma}_i^X$ denote the sample autocovariance matrices (similar to the true autocovariances Γ_i^X defined in Section 5). The estimation algorithm is given by Algorithm 2.

7.3. Model Checking

Ideally we would like to know whether the various model assumptions we make in this paper, most importantly the one that the entries of B can in fact be interpreted causally, are appropriate. Obviously, this is impossible to answer just based on the observed sample of X. Nonetheless one can check these assumptions to the extent they imply testable properties of X.

For instance, to check (to a limited extent) the assumptions underlying Theorems 1 and 2 and Algorithm 1, i.e., the general statistical and causal model assumptions from Sections 4.1 and 4.2 together with A1, A2 and G1 from Section 6.1, we propose the following two tests: First, test whether $R_t(\hat{U}_1, \hat{U}_2)$ is independent of $(X_{t-2-j})_{j=0}^{J}$, for $(\hat{U}_1, \hat{U}_2)$ as defined in Algorithm 2, and for say $J = 2$. (If Algorithm 2 finds no $(\hat{U}_1, \hat{U}_2)$ then the test is already failed.) Second, check whether all components of X_t are non-Gaussian using e.g. the Kolmogorov-Smirnov test (Conover, 1971) for Gaussianity.

Note that under the mentioned assumptions, both properties of X do in fact hold true. Regarding the independence statement, this follows from Lemmas 4 and 2. W.r.t. the non-Gaussianity statement, this follows from the fact (Ramachandran, 1967, Theorem 7.8) that the distribution of an infinite weighted sum of non-Gaussian random variables is again non-Gaussian. It should be mentioned that the first test can also be used to check (to a limited extent) the assumptions underlying Theorem 3 and Algorithm 2.

8. Experiments

In this section we evaluate the two algorithms proposed in Section 7 on synthetic and real-world data and compare them to the practical Granger causation estimator. Keep in mind that the latter is defined by replacing the covariances in equation (3) by sample covariances.

8.1. Synthetic Data

We empirically study the behavior of Algorithms 1 and 2 on simulated data, in dependence on the sample length. Note that, based on theoretical considerations (see Section 4.3), it can be expected that the error of the practical Granger estimator is substantially bounded away from zero in the generic case.

8.1.1. Algorithm 1

Here we evaluate Algorithm 1.

Experimental setup: We consider the case of a 2-variate X and a 1-variate Z, i.e., $K_X = 2, K_Z = 1$. We consider sample lengths $L = 100, 500, 1000, 5000$ and for each sample length we do 20 runs. In each run we draw the matrix A uniformly at random from the stable matrices and then randomly draw a sample of length L from a VAR process $W = (X, Z)^\top$ with A as transition matrix and noise N_t^k distributed according to a super-Gaussian mixtures of Gaussians. Then we apply Algorithm 1 and the practical Granger causation estimator on the sample of *only* X.

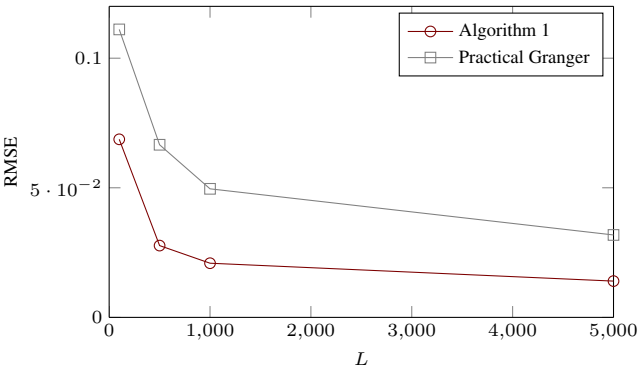

Figure 1. RMSE of Algorithm 1 and the practical Granger estimator as a function of sample length L.

Outcome: We calculated the root-mean-square error (RMSE) of Algorithm 1, i.e., $\frac{1}{20}\sum_{n=1}^{20}(B_n^{\text{est}} - B_n^{\text{true}})^2$, where $B_n^{\text{est}}, B_n^{\text{true}}$ denotes the output of Algorithm 1 and the true B, respectively, for each run n. The RMSE as a function of the sample length L is depicted in Figure 1, along with the RMSE of the practical Granger algorithm.

Discussion: This suggests that for $L \to \infty$ the error of Algorithm 1 is negligible, although it may not converge to zero. The error of the practical Granger estimator for $L \to \infty$ is still small but substantially bigger than that of Algorithm 1.

8.1.2. Algorithm 2

Here we empirically establish the error of Algorithm 2, more precisely the deviation between the true B and the best out of the several estimates that Algorithm 2 outputs. Obviously in general it is unknown which of the outputs of Algorithm 2 is the best estimate. However here we rather want to establish that asymptotically, the output of Algorithm 2 in fact contains the true B. Also we compare Algorithm 2 to the practical Granger estimator, although it needs to be said, that the latter is usually not applied to univariate time series.

Experimental setup: We consider the case of 1-variate X and Z, i.e., $K_X = K_Z = 1$. We consider sample lengths $L = 10^1, 10^2, \ldots, 10^7$ and for each sample length we do 20 runs. In each run we draw the matrix A uniformly at random from the stable matrices with the constraint that the lower left entry is zero and then randomly draw a sample of length L from a VAR process $W = (X, Z)^\top$ with A as transition matrix and standard normally distributed noise N. Then we apply Algorithm 2 and the practical Granger

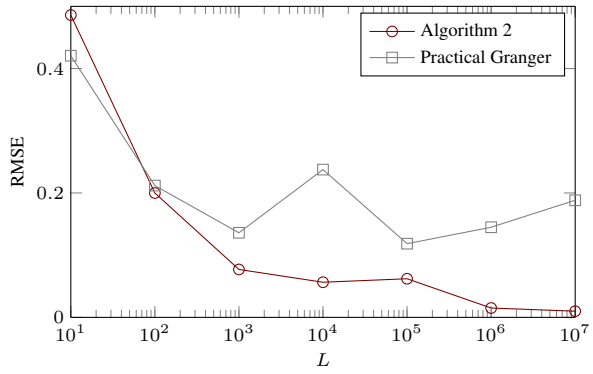

Figure 2. RMSE of Algorithm 2 and the practical Granger estimator as a function of sample length L.

causation estimator on the sample of only X.

Outcome: We calculated the root-mean-square error (RMSE) of Algorithm 2, i.e., $\frac{1}{20}\sum_{n=1}^{20}(B_n^{\text{best est}} - B_n^{\text{true}})^2$, where $B_n^{\text{best est}}, B_n^{\text{true}}$ denotes the best estimate for B returned by Algorithm 2 (i.e., the one out of the two outputs that minimizes the RMSE) and true B for each run n, respectively. The RMSE as a function of the sample length L is depicted in Figure 2, along with the RMSE of the practical Granger estimator.

Discussion: This empirically shows that the set of two outputs of Algorithm 2 asymptotically seem to contain the true B. However, it takes at least 1000 samples to output reasonable estimates. As expected, the practical Granger estimator does not seem to converge against the true B.

8.2. Real-World Data

Here we examine how Algorithm 1 performs on a real-world data set.

Experimental setup: We consider a time series Y of length 340 and the three components: cheese price Y^1, butter price Y^2, milk price Y^3 (recorded monthly from January 1986 to April 2014, `http://future.aae.wisc.edu/tab/prices.html`). We used the following estimators: We applied practical Granger estimation to the full time series Y (i.e., considering $X = Y$) and denote the outcome by A_{fG}. We applied practical Granger estimation to the reduced time series $(Y^1, Y^2)^\top$ (i.e., considering $X = (Y^1, Y^2)^\top$) and denote the outcome by B_{pG}. We applied Algorithm 1 to the full time series Y (i.e., considering $X = Y$), while assuming an additional hidden univariate Z, and denote the outcome by $\bar{A}_{\text{fA}}$. We applied Algorithm 1 to the reduced time series $(Y^1, Y^2)^\top$ (i.e., considering $X = (Y^1, Y^2)^\top$), while assuming an additional hidden univariate Z, and denote the outcome by $\tilde{A}_{\text{pA}}$. Furthermore we do a model check as suggested in Section 7.3, although the sample size may be too small for the independence test to work reliably.

Outcome: The outputs are:

$$A_{\text{fG}} = \begin{pmatrix} 0.8381 & 0.0810 & 0.0375 \\ 0.0184 & 0.9592 & -0.0473 \\ 0.2318 & 0.0522 & 0.7446 \end{pmatrix},$$

$$B_{\text{pG}} = \begin{pmatrix} 0.8707 & 0.0837 \\ -0.0227 & 0.9559 \end{pmatrix},$$

$$\bar{A}_{\text{fA}} = \begin{pmatrix} 0.8809 & 0.1812 & 0.1016 & -0.1595 \\ 0.0221 & 1.0142 & -0.0290 & -0.0492 \\ 0.2296 & 0.1291 & 0.8172 & -0.1143 \\ 1.0761 & 0.6029 & -0.7184 & 0.4226 \end{pmatrix},$$

$$\tilde{A}_{\text{pA}} = \begin{pmatrix} 0.9166 & 0.0513 & -0.0067 \\ -0.0094 & 0.9828 & -0.0047 \\ -0.0031 & 0.1441 & -0.2365 \end{pmatrix}.$$

The outcome of the model check, based on a significance level of 5%, is the following: the hypothesis of Gaussianity is rejected. Also the independence hypothesis stated in Section 7.3 is rejected. The latter implies that the model assumptions underlying Algorithm 1 are probably wrong.

Discussion: We consider A_{fG} as ground truth. Intuitively, non-zero entries at positions $(i, 3)$ can be explained by the milk price influencing cheese/butter prices via production costs, while non-zero entries at positions $(3, j)$ can be explained by cheese/butter prices driving the milk price via demand for milk. The explanation of non-zero entries at positions (1, 2) an (2, 1) is less clear. One can see that the upper left 2×2 submatrix of $\tilde{A}_{\text{pA}}$ is quite close to that of A_{fG} (the RMSE over all entries is 0.0753), which shows that Algorithm 1 works well in this respect. Note that B_{pG} is even a bit closer (the RMSE is 0.0662). However, the upper right 2×1 matrix of $\tilde{A}_{\text{pA}}$ is not close to a scaled version of the upper right 2×1 submatrix of A_{fG} (which corresponds to C). This is in contrast to what one could expect based on Theorem 2. $\bar{A}_{\text{fA}}$ can be seen as an alternative ground truth. It is important to mention that the estimated order (lag length) of the full time series Y is 3, according to Schwarz's criterion (SC) (Lütkepohl, 2006), which would violate our assumption of a VAR process of order 1 (Section 4.1). The model check seems to detect this violation of the model assumptions.

9. Conclusions

We considered the problem of causal inference from observational time series data. Our approach consisted of two parts: First, we examined possible conditions for identifiability of causal properties of the underlying system from the given data. Second, we proposed two estimation algorithms and showed that they work on simulated data under the respective conditions from the first part.

Acknowledgements

Kun Zhang was supported in part by DARPA grant No. W911NF-12-1-0034.

References

Besserve, M., Schölkopf, B., Logothetis, N. K., and Panzeri, S. Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *Journal of Computational Neuroscience*, 29(3):547–566, 2010. doi: 10.1007/s10827-010-0236-5. URL http://dx.doi.org/10.1007/s10827-010-0236-5.

Boyen, X., Friedman, N., and Koller, D. Discovering the hidden structure of complex dynamic systems. In *Proceedings of the 15 th Conference on Uncertainty in Artificial Intelligence*, pp. 91–100. Morgan Kaufmann, San Francisco, 1999.

Chu, T. and Glymour, C. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9:967–991, 2008.

Conover, W.J. *Practical Nonparametric Statistics*. John Wiley & Sons, 1971.

Eichler, M. Causal inference in time series analysis. In Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.), *Causality*, pp. 327–354. John Wiley and Sons, Ltd, 2012.

Etesami, J., Kiyavash, N., and Coleman, T.P. Learning minimal latent directed information trees. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2726–2730, 2012.

Geiger, P., Zhang, K., Gong, M., Janzing, D., and Schölkopf, B. Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components. *ArXiv e-prints*, 2015. arXiv:1411.3972 [stat.ML].

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969. ISSN 00129682.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362 – 378, 2008. doi: http://dx.doi.org/10.1016/j.ijar.2008.02.006.

Hyvaerinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, (11):1709–1731, 2010.

J. E. Dennis, Jr., Traub, J. F., and Weber, R. P. The algebraic theory of matrix polynomials. *SIAM Journal on Numerical Analysis*, 13(6):831–845, 1976.

Jalali, A. and Sanghavi, S. Learning the dependence graph of time series with latent factors. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Kagan, A. M., Linnik, Y. V., and Rao, C. R. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg, New York, 2006.

Oh, S. M., Ranganathan, A., Rehg, J. M., and Dellaert, F. A variational inference method for switching linear dynamic systems. Technical report, 2005.

Pearl, J. *Causality*. Cambridge University Press, 2000.

Ramachandran, B. *Advanced theory of characteristic functions*. Series in probability and statistics. Statistical Pub. Society, 1967.

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

Schreiber, T. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search*. MIT, Cambridge, MA, 2nd edition, 2000.