The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Vector Autoregressive Models with Latent Processes

Saber Salehkaleybar,[*] Jalal Etesami,[*†] Negar Kiyavash,[†‡] Kun Zhang[◊]
[*]Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, USA.
[†]Department of ISE, University of Illinois at Urbana-Champaign Urbana, USA.
[‡]Department of ECE, University of Illinois at Urbana-Champaign, Urbana, USA.
[◊]Department of Philosophy, Carnegie Mellon University, Pittsburgh, USA.
{sabersk,etesami2,kiyavash}@illinois.edu,kunz1@cmu.edu

Abstract

We study the problem of learning the support of transition matrix between random processes in a Vector Autoregressive (VAR) model from samples when a subset of the processes are latent. It is well known that ignoring the effect of the latent processes may lead to very different estimates of the influences among observed processes, and we are concerned with identifying the influences among the observed processes, those between the latent ones, and those from the latent to the observed ones. We show that the support of transition matrix among the observed processes and lengths of all latent paths between any two observed processes can be identified successfully under some conditions on the VAR model. From the lengths of latent paths, we reconstruct the latent subgraph (representing the influences among the latent processes) with a minimum number of variables uniquely if its topology is a directed tree. Furthermore, we propose an algorithm that finds all possible minimal latent graphs under some conditions on the lengths of latent paths. Our results apply to both non-Gaussian and Gaussian cases, and experimental results on various synthetic and real-world datasets validate our theoretical results.

Introduction

Identifying causal influences among time series is a problem of interest in many fields. In macroeconomics, for instance, researchers seek to understand what factors contribute to economic fluctuations and how they interact with each other (Lütkepohl and Krätzig 2004). In neuroscience, many researchers focus on learning the interactions between different regions of brain by analyzing neural spike trains (Roebroeck, Formisano, and Goebel 2005; Kim et al. 2014).

Granger causality (Granger 1969), transfer entropy (Schreiber 2000), and directed information (Massey 1990; Marko 1973; Quinn, Kiyavash, and Coleman 2013; Etesami and Kiyavash 2014) are some of the most commonly used measures in the literature to calculate time-delayed dependence structures in time series. Measuring the reduction of uncertainty in one variable after observing another variable is the key concept behind such measures. Under certain assumptions, these measures may represent causal relations among the variables (Pearl 2009; Spirtes, Glymour, and Scheines 2000). In (Eichler 2012), an overview of various definitions of causation is given for time series.

In this work, we study the causal identification problem in VAR models when only a subset of times series is observed. More precisely, we assume that the available measurements are a set of random processes $\vec{X}(t) \in \mathbb{R}^n$ which, together with another set of latent random processes $\vec{Z}(t) \in \mathbb{R}^m$, where $m \leq n$ form a first order VAR model as follows:

$$\begin{bmatrix} \vec{X}(t+1) \\ \vec{Z}(t+1) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \vec{X}(t) \\ \vec{Z}(t) \end{bmatrix} + \begin{bmatrix} \vec{\omega}_X(t+1) \\ \vec{\omega}_Z(t+1) \end{bmatrix}. \quad (1)$$

Here we assume that observed data were measured at the right causal frequency of the VAR process; otherwise one may need to consider the effect of the sampling procedure such as subsampling or temporal aggregation (Danks and Plis 2013; Gong et al. 2015; 2017). Under certain assumptions (e.g., causal sufficiency), the support of the transition matrix corresponds to the causal structure between these processes (Granger 1969; Spirtes, Glymour, and Scheines 2000; Pearl 2009). If we ignore the influence of latent processes and just regress $\vec{X}(t+1)$ on $\vec{X}(t)$, we may get a wrong estimate of the transition matrix between observed processes (see the example in (Geiger et al. 2015)). Hence, it is crucial to consider the presence of latent processes and their influences on the observed processes.

Contributions: The contributions of this paper are as follows: we propose a learning approach that recovers the *observed sub-network* (support of A_{11}) by regressing the observed vector $\vec{X}(t+1)$ on a set of its past observations (not just $\vec{X}(t)$) as long as the graph representation of *latent subnetwork* (support of A_{22}) is a directed acyclic graph (DAG)[1]. We also derive a set of sufficient conditions under which we can uniquely recover the influences from latent to observed processes, (support of A_{12}) and also the influences among the latent variables, (support of A_{22}). Additionally, we propose a sufficient condition under which the support of the complete transition matrix can be recovered uniquely.

More specifically, we show that under an assumption on the observed to latent noise power ratio, if neither of the sub-

[1]Support of matrix A is a matrix with the same size where entry (i, j) is equal to one if the corresponding entry in A is nonzero. Otherwise, it would be zero.

matrices A_{12} and A_{21} are zero, it is possible to determine the length of all directed *latent paths*[2]. We refer to this information as *linear measurements*[3]. This information reveals important properties of the causal structure among the latent and observed processes, i.e., support of $[0, A_{12}; A_{21}, A_{22}]$. We call this sub-network of a VAR model *unobserved network*. We show that in the case that the unobserved network is a directed tree and each latent variable has at least two parents and two children, a straightforward application of (Patrinos and Hakimi 1972) can recover the unobserved network uniquely. Furthermore, we propose Algorithm 1 that recovers the support of A_{22} and A_{12} given the linear measurements when only the latent sub-network is a directed tree plus some extra structural assumptions (see Assumption 2). Lastly, we study the causal structures of VAR models in a more general case in which there exists at most one directed latent path of length $k \geq 2$ between any two observed processes (see Assumption 3). For such VAR models, we propose Algorithm 2 that can recover all possible unobserved networks with minimum number of latent processes. Our results apply to both non-Gaussian and Gaussian cases, and experimental results on various synthetic and real-world datasets validate our theoretical results. All proofs can be found in supplemental material.

Related works: The problem of recovering latent causal structure for time series has been studied in the literature. Assuming that connections between observed variables are sparse and each latent variable interacts with many observed variables, it has been shown that the transition matrix between observed variables can be identified in a VAR model (Jalali and Sanghavi 2012). However, their approach focuses on learning only the observed sub-network. (Boyen, Friedman, and Koller 1999) applied a method based on expectation maximization (EM) to infer properties of partially observed Markov processes, without providing theoretical analysis for identifiability. (Geiger et al. 2015) showed that if the exogenous noises are independent non-Gaussian and additional so-called genericity assumptions hold, then the sub-networks A_{11} and a part of A_{12} are uniquely identifiable. However, these assumptions may not hold true in a real-world dataset even with three variables (Geiger et al. 2015). They also presented a result in which they allowed Gaussian noises in their VAR model and obtained a set of conditions under which they can recover up to $\binom{2n}{n}$ candidate matrices for A_{11}. Their learning approach is also based on EM and approximately maximizes the likelihood of a parametric VAR model with a mixture of Gaussians as noise distribution. Recently, (Etesami, Kiyavash, and Coleman 2016) studied a network of processes (not necessary a VAR model) whose underlying structure is a polytree and introduced an algorithm that can learn the entire casual structure (observed and unobserved networks) using a particular discrepancy measure.

Compared to related works, we show the identifiability of new class of structures in the presence of latent processes. Unlike (Geiger et al. 2015), we do not assume the non-Gaussian distribution of the exogenous noises or those genericity assumptions. Moreover, our results do not rely on the assumption that connections between observed variables are sparse or each latent variables interacts with many observed variables as in (Jalali and Sanghavi 2012). Furthermore, these works (Geiger et al. 2015; Jalali and Sanghavi 2012) can uniquely identify at most a part of transition matrix (A_{11} or a part of A_{12}). We should emphasize that our proposed methods are not necessarily more general than existing ones, but provide correct solutions under sensible assumptions with graphical interpretation.

[2]A directed path is a latent path if it connects two observed variables and all the intermediate variables on that path are latent.

[3]This is because it can be inferred from the observational data using linear regression.

Problem Definition

In this part, we review some basic definitions and our notation. Throughout this paper, we use an arrow over the letters to denote vectors. We assume that the time series are stationary and denote the autocorrelation of $\vec{X}$ by $\gamma_X(k) := \mathbb{E}[\vec{X}(t)\vec{X}(t-k)^T]$. We denote the support of a matrix A by $Supp(A)$ and use $Supp(A) \subseteq Supp(B)$ to indicate $[A]_{ij} = 0$ whenever $[B]_{ij} = 0$. We also denote the Fourier transform of g by $\mathcal{F}(g)$ and it is given by $\sum_{h=-\infty}^{\infty} g(h)e^{-h\Omega j}$.

In a directed graph $G = (V, \overrightarrow{E})$ with the node set V and the edge set $\overrightarrow{E}$, we denote the set of parents of a node v by $\mathcal{P}_v := \{u : (u, v) \in \overrightarrow{E}\}$ and the set of its children by $\mathcal{C}_v := \{u : (v, u) \in \overrightarrow{E}\}$. The skeleton of a directed graph G is the undirected graph obtained by removing all the directions in G.

System Model

Consider the VAR model in (1). Let $\vec{\omega}_X(t) \in \mathbb{R}^n$ and $\vec{\omega}_Z(t) \in \mathbb{R}^m$ be i.i.d random vectors with mean zero. For simplicity, we denote the matrix $[A_{11}, A_{12}; A_{21}, A_{22}]$ by A. Our goal is to recover $Supp(A)$ from observational data, i.e., $\{\vec{X}(t)\}$. Rewrite (1) as follows

$$\vec{X}(t+1) = \sum_{k=0}^{t} A_k^* \vec{X}(t-k) + A_{12}A_{22}^t \vec{Z}(0) + \sum_{k=0}^{t-1} \tilde{A}_k \vec{\omega}_Z(t-k) + \vec{\omega}_X(t+1), \quad (2)$$

where $A_0^* := A_{11}$, $A_k^* := A_{12}A_{22}^{k-1}A_{21}$ for $k \geq 1$, and $\tilde{A}_k := A_{12}A_{22}^k$.

Assumption 1. *We assume that the A_{22} is acyclic, i.e., $\exists\, 0 < l \leq m$, such that $A_{22}^l = 0$.*

Based on the above assumption, for $t \geq l$, Equation (2) becomes[4]

$$\vec{X}(t+1) = \sum_{k=0}^{l} A_k^* \vec{X}(t-k) + \sum_{k=0}^{l-1} \tilde{A}_k \vec{\omega}_Z(t-k) + \vec{\omega}_X(t+1). \quad (3)$$

[4]Note that the limits of summations in (3) are changed.

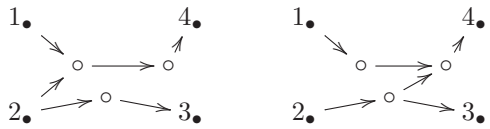


Figure 1: Two unobserved networks with the same linear measurements. White circles denote latent nodes.

We are interested in recovering the set $\{Supp(A_k^*)\}_{k=0}^l$ because it captures important information about the structure of the VAR model. Specifically, $Supp(A_0^*) = Supp(A_{11})$; so it represents the direct causal influences between the observed variables and $Supp(A_k^*)$ for $k \geq 1$ determines whether at least one directed path of length $k+1$ exists between any two observed nodes which goes through the latent sub-network[5]. We will make use of this information in our recovery algorithm. We call the set of matrices $\{Supp(A_k^*)\}_{k\geq 0}$, *linear measurements*. In Section 4, we present a set of sufficient conditions under which given the linear measurements, we can recover the entire or most parts of the unobserved network uniquely.

Note that in general, the linear measurements cannot uniquely specify the unobserved network. For example, Figure 1 illustrates two different unobserved networks that both share the same set of linear measurements, $A_k^* = 0$ for $k > 2$ and the only nonzero entries of A_1^* and A_2^* are $\{(3,2)\}$ and $\{(4,1),(4,2)\}$, respectively [6].

Identifiability of the Linear Measurements

As we need the linear measurements for our structure learning, in this section, we study a sufficient condition under which we can recover the linear measurements from the observed processes $\{\vec{X}(t)\}$. To do so, we start off by rewriting Equation (3) as follows,

$$\vec{X}(t+1) = \mathcal{A}\vec{\mathcal{X}}_{t-l:t} + \sum_{k=0}^{l-1} \tilde{A}_k \vec{\omega}_Z(t-k) + \vec{\omega}_X(t+1), \quad (4)$$

where $\mathcal{A} := [A_0^*, ..., A_l^*]$, and $\vec{\mathcal{X}}_{t-l:t} := [\vec{X}(t); \cdots; \vec{X}(t-l)]$. By projecting $\tilde{A}_k\vec{\omega}_Z(t-k)$ onto the vector space spanned by the observed processes, i.e., $\{\vec{X}(t), ..., \vec{X}(t-l)\}$, we obtain

$$\tilde{A}_k\vec{\omega}_Z(t-k) = \sum_{r=0}^{l} C_r^s \vec{X}(t-r) + \vec{N}_Z(t-k),\ 0 \leq k \leq l-1, \quad (5)$$

where $\{\vec{N}_Z(t-k)\}$ denote the residual terms and $\{C_r^s\}$ are the corresponding coefficient matrices. Substituting (5) into (4) implies

$$\vec{X}(t+1) = \mathcal{B}\vec{\mathcal{X}}_{t-l:t} + \vec{\theta}(t+1), \quad (6)$$

where $\mathcal{B} := [B_0^*, ..., B_l^*]$, $B_k^* := A_k^* + \sum_{s=0}^{l-1} C_k^s$, and $\vec{\theta}(t+1) := \vec{\omega}_X(t+1) + \sum_{k=0}^{l-1} \vec{N}_Z(t-k)$. Note that by this representation, $\vec{\theta}(t+1)$ is orthogonal to $\vec{\mathcal{X}}_{t-l:t}$. Hence, Equation (6) shows that the minimum mean square error (MMSE) estimator can learn the coeffiecient matrix $\mathcal{B}$ given the observed processes. More precisely, let $\Gamma_X(l) := \mathbb{E}\{\vec{\mathcal{X}}_{t-l:t}\vec{\mathcal{X}}_{t-l:t}^T\}$, then we have

$$\mathcal{B} = [\gamma_X(1), .., \gamma_X(l+1)] \times \Gamma_X(l)^{-1}. \quad (7)$$

Proposition 1. *Under Assumption 1, for the stationary VAR model in (1), we have*

$$||B_k^* - A_k^*||_1 \leq \sqrt{n(l-k-1)M/L}||A_{12}||_2||A_{22}||_2^{k+1},$$

where $M := \lambda_{max}(\Gamma_{\omega_Z}(0))$ *and* $L := \lambda_{min}(\Gamma_X(0))$.

This result implies that we can asymptotically recover the support of $\{A_k^*\}_{k=0}^l$ as long as the absolute values of non-zero entries of A_k^* are bounded away from zero by $2\sqrt{n(l-k-1)\frac{M}{L}}||A_{12}||_2||A_{22}||_2^{k+1}$. In Appendix (the second section), we explained how these bounds can be estimated from observational data.

Proposition 2. *Let* $\Sigma_X = \sigma_X^2 I_{n\times n}$ *and* $\Sigma_Z = \sigma_Z^2 I_{m\times m}$ *be the autocovariance matrices of* $\vec{\omega}_X(t)$ *and* $\vec{\omega}_Z(t)$*, respectively. Then, the ratio* M/L *strictly increases by decreasing* σ_X^2/σ_Z^2.

Proposition 2 implies that when the σ_X^2/σ_Z^2 increases, M/L will decrease, and based on the bound in Proposition 1, the estimation error will decrease (it goes to zero as σ_X^2/σ_Z^2 tends to infinity). This shows that recovering the linear measurements is much easier in high σ_X^2/σ_Z^2 regime as illustrated in Figure 3b. Note that Proposition 1 stresses a suffiecient condition for recovering the linear measurements. As shown in Figure 3b, in practice, the actual estimation error is much smaller than the bound in Proposition 1. In the next section, we will make use of $\{Supp(A_k^*)\}_{k>0}$ to recover the unobserved network. We assume that the correct linear measurements can be obtained from matrix $\mathcal{B}$.

In order to estimate the support of matrix $\mathcal{B}$ from a finite number of samples drawn from the observed processes, say $\{\vec{X}(t)\}_{t=1}^T$, first we obtain the lag length l in (6) by AIC or FPE criterion (see Chapter 4 in (Lütkepohl 2005)). Afterwards, we can estimate the coefficient matrix $\mathcal{B}$, using an empirical estimator for $\Gamma_X(l)$, $\{\gamma_X(h)\}_{h=1}^{l+1}$, and then applying (7). Denote the result of this estimation by $\mathcal{B}_T$. It can be shown that (Lütkepohl 2005), $\sqrt{T}\text{vec}(\mathcal{B}_T - \mathcal{B}) \xrightarrow[T\to\infty]{d} \mathcal{N}(0, \Gamma_X^{-1}(l) \otimes \Sigma)$, where $\xrightarrow{d}$ denotes convergence in distribution, and Σ is the autocovariance matrix of $\vec{\theta}(t)$. vec(.) transforms a matrix to a vector by stacking its columns and $\otimes$ is the Kronecker product. Having the estimates of $\Gamma_X(l)$ and Σ, we can test whether the entries of matrix $\mathcal{B}$ are greater than the bounds in Proposition 1 (see Chapter 3 in (Lütkepohl 2005)).

Learning the Unobserved Network

Recall that we refer to $Supp([0, A_{12}; A_{21}, A_{22}])$ as the unobserved network and $Supp(A_{22})$ as the latent sub-network.

[5]Herein, we exclude degenerate cases where there is a direct path from an observed node to another one with length k but the corresponding entry in matrix $Supp(A_k^*)$ is zero. In fact, such special cases can be resolved by small perturbation of nonzero entries in matrix A. In the causal discovery literature, this assumption is known as faithfulness (Spirtes, Glymour, and Scheines 2000).

[6]In this work, a graph is a representation of transition matrix A. In particular, there is a directed edge from node j to node i if entry (i,j) of the matrix is nonzero.

We present three algorithms that take the linear measurements $\{Supp(A_k^*)\}_{k\geq 0}$ as their input. The first algorithm recovers the entire unobserved network uniquely as long as it is a directed tree and each latent node has at least two parents and two children. The output of the second algorithm is $Supp([0, A_{12}; \widehat{A}_{21}, A_{22}])$, where $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$. This is guaranteed whenever the latent sub-network is a directed tree and some extra conditions are satisfied on how the latent and observed nodes are connected. The third algorithm finds the set of all possible networks with minimum number of latent nodes that are consistent with the measurements. This algorithm is able to do so when there exists at most one directed latent path of any arbitrarily length between two observed nodes. A directed path is latent if all the intermediate variables on that path are latent.

Unobserved Network is a Directed Tree

Authors in (Patrinos and Hakimi 1972) introduced a necessary and sufficient condition for recovering a weighted directed tree uniquely from a valid distance matrix D defined on the observed nodes,[7] and also proposed a recovery algorithm. The condition is as follows: every latent node must have at least two parents and two children. A matrix D, in (Patrinos and Hakimi 1972), is a valid distance matrix, when $[D]_{ij}$ equals the sum of all the weights of those edges that belong to the directed path from i to j, and $[D]_{ij} = 0$, if there is no directed path.

The algorithm in (Patrinos and Hakimi 1972) has two phases. In the first phase, it creates a directed graph among the observed nodes with the adjacency matrix $Supp(D)$. In the second phase, it recursively finds and removes the circuits by introducing latent nodes for each circuit[8]. For more details, see (Patrinos and Hakimi 1972).

In order to adopt (Patrinos and Hakimi 1972)'s algorithm for learning the unobserved network, we introduce a valid distance matrix using our linear measurements as follows, $D_{ij} = k + 1$ if $[Supp(A_k^*)]_{ji} \neq 0$ and 0, otherwise. Recall that $[Supp(A_k^*)]_{ji}$ indicates whether there exists a directed latent path from i to j of length $k+1$ in the unobserved network. From theorem 8 in (Patrinos and Hakimi 1972), it is easy to show that the unobserved network can be recovered uniquely from above distance matrix if its topology is a directed tree and every latent node has at least two parents and two children.

Latent Sub-network is a Directed Tree

Definition 1. *We denote the subset of observed nodes that are parents of a latent node h by $\mathcal{P}_h^O$ and denote the subset of observed nodes for which h is a parent, by $\mathcal{C}_h^O$. We further denote the set of all leaves in the latent sub-network by $\mathcal{L}$.*

We consider learning an unobserved network G that satisfies the following assumptions.

[7]The skeleton of the recovered tree is the same as the original one but not necessary the weights.

[8]In a directed graph, a circuit is a cycle after removing all the directions.

Algorithm 1 DTR Algorithm

Input: $\{Supp(A_k^*)\}_{k\geq 1}$
Find $\{l_i\}$ using (8) and set $U := \emptyset$.
Find R_i, M_i from (9) for all $1 \leq i \leq n$.
for $i = 1, ..., n$ **do**
  $Y_i := \{j : j \neq i \wedge l_j = l_i\}$
  if $\forall j \in Y_i, (R_j \not\subseteq R_i) \vee (R_j = R_i \wedge M_i \subseteq M_j)$ **then**
    if $i = \min\{k : R_k = R_i \wedge M_k = M_i\}$ **then**
      Create node h_i and set $\mathcal{P}_{h_i}=\{i\}$, $U \leftarrow \{i\} \cup U$
    end if
  end if
end for
for every latent node h_s **do**
  if $\exists h_k, (l_k = l_s + 1) \wedge (R_s \subseteq R_k)$ **then**
    $\mathcal{P}_{h_s} \leftarrow \{h_k\} \cup \mathcal{P}_{h_s}$
  end if
  $\mathcal{C}_{h_s} \leftarrow \{j : [A_1^*]_{js} \neq 0\}$
end for
for $i = 1, ..., n$ **do**
  if $\exists\, j \in U$, s.t. $M_j \subseteq M_i$ **then**
    $\mathcal{P}_{h_j} \leftarrow \{i\} \cup \mathcal{P}_{h_j}$
  end if
end for

Assumption 2. *Assume that the latent sub-network of G is a directed tree. Furthermore, for any latent node h in G, (i) $\mathcal{P}_h^O \not\subseteq \cup_{h\neq j}\mathcal{P}_j^O$ and, (ii) if h is a leaf of the latent sub-network, then $\mathcal{C}_h^O \not\subseteq \cup_{i\in\mathcal{L}, i\neq h}\mathcal{C}_i^O$.*

This assumption states that the latent sub-network of G must be a directed tree such that each latent node in G has at least one unique parent in the set of observed nodes. That is, a parent who is not shared with any other latent node. Furthermore, each latent leaf has at least one unique child among the observed nodes. For instance, when $Supp(A_{22})$ represents a directed tree and both $Supp(A_{12})$ and $Supp(A_{21})$ contain identity matrices, Assumption 2 holds. As we will see later in Experimental Results (Figure 3c), a large portion of randomly generated graphs satisfy Assumption 2.

Figure 2e illustrates a simple network that satisfies Assumption 2 in which the unique parents of latent nodes a, b, c, and d are $\{1\}$, $\{3\}$, $\{2\}$, and $\{4\}$, respectively. The unique children of latent leaves c and d are $\{5\}$ and $\{2, 4\}$, respectively.

Theorem 1. *Among all unobserved networks that are consistent with the linear measurements induced from (1), any graph G that satisfies Assumption 2 has the minimum number of latent nodes.*

Note that if Assumption 2 is violated, one can find many unobserved networks that are consistent with the linear measurements but are not minimum (in terms of the number of latent nodes). For example, the network in Figure 2a satisfies Assumption 2 (ii) but not (i). Figure 2b depicts an alternative network with the same linear measurements as the network in Figure 2a but it has fewer number of latent nodes. Similarly, the graph in Figure 2c satisfies Assumption 2 (i) but

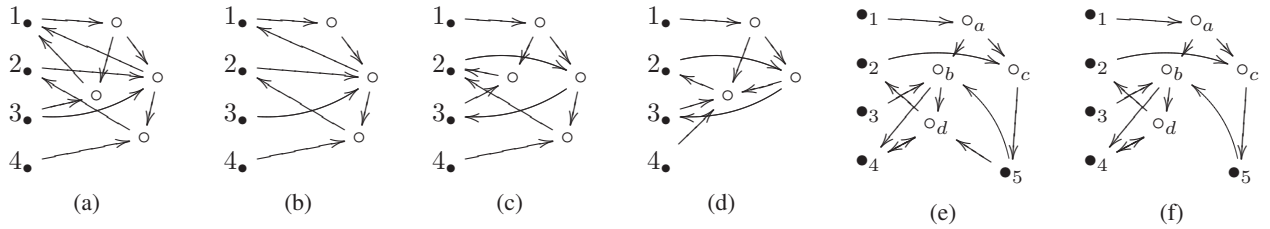


Figure 2: Latent nodes are indicated by white circles. Graph (a) satisfies (ii) but not (i) and it can be reduced to (b). Graph (c) satisfies (i) but not (ii) and it can be reduced to (d). (e) and (f) satisfy Assumption 2 and have the same induced linear measurements but $Supp(A_{21})_{(f)} \subset Supp(A_{21})_{(e)}$.

not (ii). Figure 2d shows an alternative graph with one less latent node.

Theorem 2. *Consider an unobserved network G with adjacency matrix $Supp([0, A_{12}; A_{21}, A_{22}])$. If G satisfies Assumption 2, then its corresponding linear measurements uniquely identify G upto $Supp([0, A_{12}; \widehat{A}_{21}, A_{22}])$, where $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$.*

Figure 2e gives an example of a network satisfying Assumption 2 and an alternative network, Figure 2f, with the same linear measurements which departs from the Figure 2e in the A_{21} component.

Next, we propose the directed tree recovery (DTR) algorithm that takes the linear measurements of an unobserved network G satisfying Assumption 2 and recovers G upto the limitation in Theorem 2. This algorithm consists of three main loops. Recall that Assumption 2 implies that each latent node has at least one unique observed parent. The first loop finds all the unique observed parents for each latent node (lines: 4-11). The second loop reconstructs $Supp(A_{22})$ and $Supp(A_{12})$ (lines: 12-17). And finally, the third loop constructs $Supp(\widehat{A}_{21})$ such that $Supp(A_{21}) \subseteq Supp(\widehat{A}_{21})$ (lines: 18-22).

The following lemma shows that the first loop of Algorithm 1 can find all the unique observed parents from each latent node. To present the lemma, we need the following definitions.

Definition 2. *For an observed node i, we define*

$$l_i := \max\{k : [A^*_{k-1}]_{si} \neq 0, \textit{ for some } s\}, \quad (8)$$

$$R_i := \{j : [A^*_{l_i-1}]_{ji} \neq 0\},\ M_i := \{(j, r) : [A^*_{r-1}]_{ji} \neq 0\}. \quad (9)$$

In the above equations, l_i denotes the length of longest directed latent path that connects node i to any observed node. R_i is the set of all observed nodes that can be reached by i with a directed latent path of length l_i and set M_i consists of all pairs (j, r) such that there exists a directed latent path from i to j with length r.

Lemma 1. *Under Assumption 2, an observed node i is the unique parent of a latent node if and only if for any other observed node j s.t. $l_i = l_j$, we have $(R_j \not\subseteq R_i) \vee (R_j = R_i \wedge M_i \subseteq M_j)$.*

In the first loop, if there exist multiple unique parents of a latent node (for instance, node 2 and node 3 in Figure 2b), we pick the one with a minimum index (lines: 7-9).

Algorithm 2 NM Algorithm

1: **Initialization:** Construct graph G_0.
2: $\mathcal{G}_0 := G_0, \mathcal{G}_s := \emptyset, \forall s > 0$
3: $k := 0$
4: **while** $\mathcal{G}_k \neq \emptyset$ **do**
5: **for** $G \in \mathcal{G}_k$ **do**
6: **for** $i', j' \in G$ **do**
7: **if** Check(G, i', j') **then**
8: $\mathcal{G}_{k+1} := \mathcal{G}_{k+1} \cup$ Merge(G, i', j').
9: **end if**
10: **end for**
11: **end for**
12: $k := k + 1$
13: **end while**
14: **Output:** $\mathcal{G}_{out} := \mathcal{G}_{k-1}$

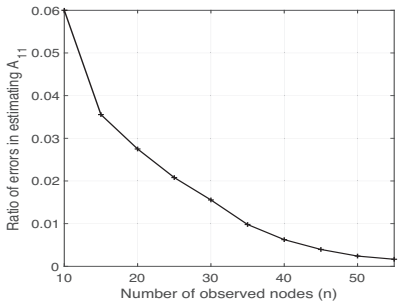
The second loop recovers $Supp(A_{22})$ based on the following observation. If a latent node h_k is the parent of latent node h_s, then h_k can reach all the observed nodes in R_s, i.e., $R_s \subseteq R_k$ and $l_k = l_s + 1$ (line: 13). Furthermore, $Supp(A_{12})$ can be recovered using the fact that an observed node j is a children of a latent node h_s, if a unique parent of h_s, e.g., s, can reach j by a directed latent path of length 2 (line: 16). Finally, the third loop reconstructs $Supp(\widehat{A}_{21})$ by adding an observed node i to the parent set of latent node h_j, if i can reach all the observed nodes that a unique parent of h_j, e.g., j, reaches (lines: 18-22).

Proposition 3. *Suppose network G satisfies Assumption 2. Then given its corresponding linear measurements, Algorithm 1 recovers G upto the limitation in Theorem 2.*

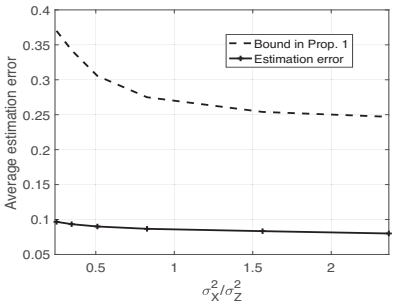
Learning More General Unobserved Networks with Minimum Number of Latent Nodes

In general, the latent sub-network may not be a tree or there may not be a unique minimal unobserved network consistent with the linear measurements (see Figure 1). Hence, we try to find an efficient approach for recovering all possible minimal unobserved networks under some conditions. In fact, without any extra conditions, finding a minimal unobserved network is NP-hard.

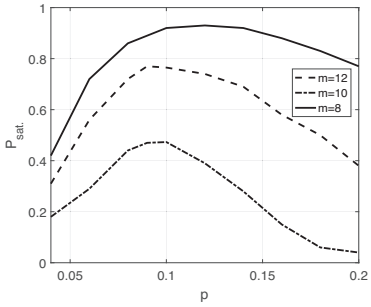
Theorem 3. *Finding an unobserved network that is both consistent with a given linear measurements and has a minimum number of latent nodes is NP-hard.*



(a) The average normalized error versus number of observed nodes.



(b) The average of estimation error versus OLNR.



(c) The probability $P_{sat.}$ versus the parameter p.

Figure 3: Average error in computing linear measurements.

Below, after some definitions, we propose the Node-Merging (NM) algorithm that returns all possible unobserved networks with minimum number of latent nodes under the following assumption.

Assumption 3. *Assume that there exists at most one directed latent path of each length between any two observed nodes.*

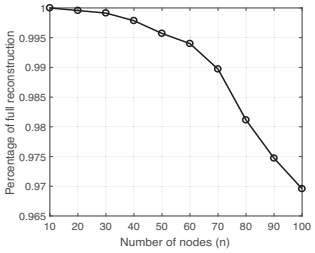
For example, the graph in Figure 2f satisfies this assumption but not the one in Figure 2e. This is because there are two directed latent paths of length 2 from node 5 to node 4.

Definition 3. *(Merging) We define merging two nodes i' and j' in graph G as follows: remove node j' and the edges between i' and j', and then give all the parents and children of $j'$to i'. We denote the resulting graph after merging i' and j' by Merge(G, i', j'). We say that two nodes i' and j' are mergeable if Merge(G, i', j') is consistent with the linear measurements of G.*

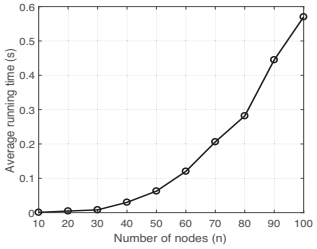
Definition 4. *(Connectedness) Consider an undirected graph $\bar{G}$ over the observed nodes which is constructed as follows: there is an edge between two nodes i and j in $\bar{G}$, if there exists $k \geq 1$ s.t. $Supp([A_k^*]_{ij}) = 1$ or $Supp([A_k^*]_{ji}) = 1$; We say that two observed nodes i and j are "connected" if there exists a path between them in $\bar{G}$.*

It can be seen that if pairs i, j and j, k are connected then node i, k are also connected. We then define a *connected class* as a subset of observed nodes in which any two nodes are connected.

Initialization: We first find the set of all connected classes, say $S_1, S_2, ..., S_C$. For each class S_c, we create a directed graph $G_{0,c}$ that is consistent with the linear measurements. To do so, for any two observed nodes $i, j \in S_c$, if $[A_r^*]_{ji} \neq 0$, we construct a directed path with length $r+1$ from node i to node j by adding r new latent nodes to $G_{0,c}$.
Merger: In this phase, for any $G_{0,c}$ from the initialization phase, we merge its latent nodes iteratively until no further latent pairs can be merged. Since the order of mergers leads to different networks with minimum number of latent nodes, the output of this phase will be the set of all such networks. Algorithm 2 summarizes the steps of NM algorithm. In this algorithm, subroutine Check(G, i', j') checks whether two nodes i' and j' are mergeable.



(a) The percentage of instances that can be reconstructed efficiently in time.



(b) Average run time of the algorithm.

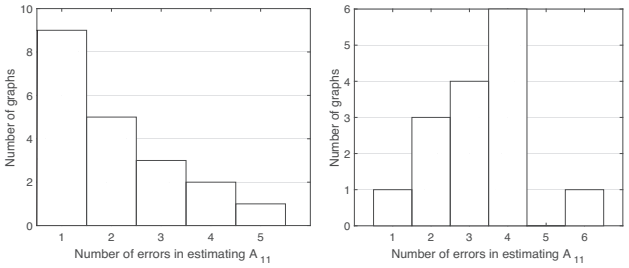
Figure 4: Recovering the minimal unobserved network for instances of DRG$(1/(2n), 1/(2n))$ where $n \in \{10, ..., 100\}$, $m = n/2$.

Theorem 4. *Under Assumptions 1 and 3, the NM algorithm returns the set of all networks that are consistent with the linear measurements and have minimum number of latent nodes.*

Experimental Results

Synthetic Data: We considered a directed random graph, denoted by DRG(p, q), such that there exists a directed link between an observed and latent node with probability p, independently across all pairs, and there is a directed link between two latent nodes with probability q. If there is a link between two nodes, we set the weight of that link uniformly from $[-a, a]$.

We utilize the method described in Section 3 to estimate linear measurements with a significance level of 0.05. In order to evaluate how well we can estimate the linear measurements, we generated 1000 instances of DRG$(0.4, 0.4)$ with $n + m = 100$, $\Sigma_X = 0.1 I_{n\times n}, \Sigma_Z = 0.1 I_{m\times m}$, and $a = 0.1$. The length of the time series was set to $T = 1000$. Let $Supp(\hat{A}_{11})$ be the estimate of support



(a) High power (b) Low power

Figure 5: Histogram of $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2$.

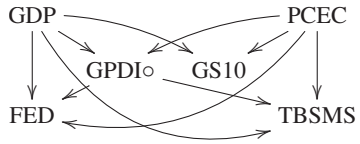


Figure 6: US macroeconomic data.

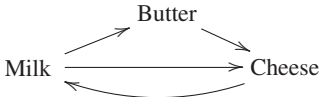


Figure 7: Dairy prices

of A_{11}. In Figure 3a, the expected estimation error, i.e. $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2/n^2$, is computed, where $||.||_F$ is the Frobenius norm. One can see that the estimation error decreases as the number of observed variables increases.

We also studied the effect of the observed to latent noise power ratio (OLNR), σ_X^2/σ_Z^2, on $||B_0^* - A_0^*||_1$, and compared it with the bound given in Proposition 1. We generated 1000 instances of DRG$(0.05, 0.05)$ with $n = 5$, $m = 5$, and $a = 0.1$. As it can be seen in Figure 3b, the average estimation error decreases as OLNR increases, as expected from Proposition 2.

We investigated what percentage of instances of the random graphs satisfy Assumption 2. We generated 1000 instances of DRG$(p, 1/n)$ with $n = 100$, and $p \in [0.04, 0.2]$. In Figure 3c, the probability of satisfying Assumption 2, $P_{sat.}$, is depicted versus p for different numbers of latent variables in the VAR model. For larger m, it is less likely to see a unique observed parent for each latent node and thus $P_{sat.}$ decreases. For a fixed m, the same phenomenon will occur if we increase p when p is relatively large. Furthermore, for small p, there might exist some latent nodes that have no observed parent or no observed children.

We also evaluated the performance of the NM algorithm in random graphs. We generated 1000 instances of DRG$(1/2n, 1/2n)$ with $n = 10, ..., 100$ and $m = n/2$, and computed the linear measurements. To save time, if for a class of connected nodes the number of latent nodes generated in the initial phase exceeds 40, we supposed that the corresponding instance cannot be recovered efficiently in time and did not proceed to the merging phase. Figures 4a and 4b depict the percentage of instances in which the algorithm can recover all possible minimal unobserved networks and the average run time (in seconds) of the algorithm, respectively[9]. This plot shows that we can recover all possible minimal unobserved networks for a large portion of instances efficiently even in relatively large networks.

US Macroeconomic Data: We considered the following set of time series from the quarterly US macroeconomic data for the period from 31-Mar-1947 to 31-Mar-2009 collected from the St. Louis Federal Reserve Economic Database (FRED) (FRE): GDP, GDPDEF, COE, HOANBS, TB3MS, PCEC, GPDI.

Assuming that the underlying dynamics is linear (see Eq. (1)), we considered the estimated VAR model over all variables as the ground truth. Then, we selected four arbitrary times series as observed processes and computed $Supp(\hat{A}_{11})$. We divided the $\binom{7}{4} = 35$ possible selections into two classes: 1) high power, where tr$(\mathbb{E}\{\omega_X(t)\omega_X(t)^T\}) > \tau$ for a fixed threshold τ; 2) low power: where tr$(\mathbb{E}\{\omega_X(t)\omega_X(t)^T\}) < \tau$. In this experiment, we set $\tau = 0.02$. In Figure 5, we plotted the histograms of $||Supp(\hat{A}_{11}) - Supp(A_{11})||_F^2$ for these two classes. As it can be seen, in the high power regime, most of the possible selections have small estimation errors.

We also considered the following six time series of US macroeconomic data during 1-Jun-2009 to 31-Dec-2016 from the same database: GDP, GPDI, PCEC, TBSMS, FEDFUND, and GS10. We obtained the causal structure among these six time series by fitting a VAR model on all of them and considered the result as our ground truth (see Figure 6). Then, we removed GPDI from the dataset and considered the remaining five time series as observe processes and checked whether the influences from the "latent" process (GPDI) can be corrected estimated. We estimated the linear measurements and gave them as an input to Algorithm 1, which successfully recovered the ground truth (the estimated structure, in which the latent process is denoted by a circle, is identical to that in Figure 6).

Dairy Prices: A collection of three US dairy prices has been observed monthly from January 1986 to December 2016 (Dai): milk, butter, and cheese prices. We estimated the VAR model on all the time series with lag length $l = 1$ and considered the resulting graph as our ground truth (see Figure 7). Next, we omitted the butter prices from the dataset and considered the milk and cheese prices as observed processes. The estimated linear measurements were: $Supp(A_0^*) = Supp(A_{11}) = [1, 1; 1, 0]$ and $Supp(A_1^*) = [0, 0; 1, 0]$. Algorithm 1 correctly recovered the true causal graph using this linear measurements. Note that so-called genericity assumptions in (Geiger et al. 2015) do not hold true for this data set (see Experiments section).

West German Macroeconomic Data: We considered the quarterly West German consumption expenditures X_1, fixed investment X_2, and disposable income X_3, during 1960-1982 (WG). Similar to the previous experiment with dairy

[9]We performed the experiment on a Mac with 2×2.4 GHz 6-Core Intel Xeon processor and 32 GB of RAM.

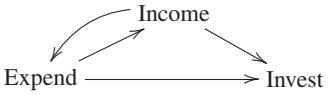


Figure 8: West German macroeconomic data.

prices, we first obtained the entire transition matrix among all the process. Figure 8 depicts the resulting graph. Next, we considered X_3 to be latent and used $\{X_1, X_2\}$ to estimate the linear measurements $Supp(A_0^*) = Supp(A_{11}) = [0, 0; 1, 1]$ and $Supp(A_1^*) = [1, 0; 1, 0]$. Using this linear measurements, Algorithm 1 recovered the true network in Figure 8 correctly.

Conclusion and Future work

We considered the problem of estimating time-delayed influence structure from partially observed time series data. Our approach consisted of two parts: First, we studied sufficient conditions under which certain aspects of the influence structure of the underlying system are identifiable. Second, we proposed two algorithms that recover the influence structures satisfying the sufficient conditions given in the first part. The proposed algorithms can construct the observed sub-network (support of A_{11}), the causal influences from latent to observed processes (support of A_{12}), and also the causal influences among the latent variables (support of A_{22}), uniquely under a set of sufficient conditions. As a future direction, we plan to extend our results to the case that A_{22} might have cycles. In this work, we have seen examples showing that unique recovery is not possible if any conditions of Assumption 2 are violated. These conditions can be a good starting point for the case that we have cycles in A_{22}.

Acknowledgments

This work was supported in part by MURI grant ARMY W911NF-15-1-0479 and ONR grant W911NF-15-1-0479.

References

Boyen, X.; Friedman, N.; and Koller, D. 1999. Discovering the hidden structure of complex dynamic systems. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 91–100. Morgan Kaufmann Publishers Inc.

(DAI) Dairy prices, dairy marketing and risk management program, University of Wisconsin. http://future.aae.wisc.edu/tab/prices.html.

Danks, D., and Plis, S. 2013. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings*.

Eichler, M. 2012. Causal inference in time series analysis. *Causality: statistical perspectives and applications* 327–354.

Etesami, J., and Kiyavash, N. 2014. Directed information graphs: A generalization of linear dynamical graphs. In *2014 American Control Conference*, 2563–2568. IEEE.

Etesami, J.; Kiyavash, N.; and Coleman, T. 2016. Learning minimal latent directed information polytrees. *Neural Computation*.

(FRE) St. Louis Federal Reserve Economic Database. http://research.stlouisfed.org/fred2/.

Geiger, P.; Zhang, K.; Gong, M.; Janzing, D.; and Schölkopf, B. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of 32th International Conference on Machine Learning (ICML 2015)*.

Gong, M.; Zhang, K.; Schölkopf, B.; Tao, D.; and Geiger, P. 2015. Discovering temporal causal relations from subsampled data. In *ICML*, 1898–1906.

Gong, M.; Zhang, K.; Schölkopf, B.; Glymour, C.; and Tao, D. 2017. Causal discovery from temporally aggregated time series. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI 17)*.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438.

Jalali, A., and Sanghavi, S. 2012. Learning the dependence graph of time series with latent factors. *ICML*.

Kim, S.; Quinn, C. J.; Kiyavash, N.; and Coleman, T. P. 2014. Dynamic and succinct statistical analysis of neuroscience data. *Proceedings of the IEEE* 102(5):683–698.

Lütkepohl, H., and Krätzig, M. 2004. *Applied time series econometrics*. Cambridge university press.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Marko, H. 1973. The bidirectional communication theory–a generalization of information theory. *Communications, IEEE Transactions on* 21(12):1345–1351.

Massey, J. 1990. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, 303–305. Citeseer.

Patrinos, A. N., and Hakimi, S. L. 1972. The distance matrix of a graph and its tree realization. *Quarterly of applied mathematics* 255–269.

Pearl, J. 2009. *Causality*. Cambridge university press.

Quinn, C. J.; Kiyavash, N.; and Coleman, T. P. 2013. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Transactions on Signal Processing* 61(12):3173–3182.

Roebroeck, A.; Formisano, E.; and Goebel, R. 2005. Mapping directed influence over the brain using granger causality and fmri. *Neuroimage* 25(1):230–242.

Schreiber, T. 2000. Measuring information transfer. *Physical review letters* 85(2):461.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

(WG) West German fixed investment, disposable income, consumption expenditures in billions of DM, 1960Q1-1982Q4. http://www.jmulti.de/data_imtsa.html.