Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding

Daniel Malinsky malinsky@jhu.edu
Johns Hopkins University
Baltimore, MD 21218, USA

Peter Spirtes ps7z@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Thuc Duy Le, Kun Zhang, Emre Kıcıman, Aapo Hyvärinen, and Lin Liu

Abstract

We present constraint-based and (hybrid) score-based algorithms for causal structure learning that estimate dynamic graphical models from multivariate time series data. In contrast to previous work, our methods allow for both "contemporaneous" causal relations and arbitrary unmeasured ("latent") processes influencing observed variables. The performance of our algorithms is investigated with simulation experiments and we briefly illustrate the proposed approach on some real data from international political economy.

Keywords: Causal Discovery, Structure Learning, Graphical Models, Time Series

1. Introduction

The FCI algorithm is a well-known constraint-based procedure for causal structure learning in settings with possible unmeasured confounding (Spirtes et al., 2000). FCI performs model selection by a sequence of conditional independence tests, and produces an estimated equivalence class of ancestral graph Markov models (Richardson and Spirtes, 2002; Zhang, 2008a,b). More recently, Ogarrio et al. (2016) introduced GFCI (Greedy FCI), which is a hybrid score-based algorithm that combines features of the Greedy Equivalence Search (GES, Chickering, 2002) with FCI. GES selects causal models by incrementally improving a model score, specifically the BIC score. GFCI executes this greedy search followed by additional conditional independence tests and orientation rules from FCI. Both of these procedures (and related variations like RFCI, FCI+) are designed for structure learning from non-temporal data. This paper extends both FCI and GFCI to the domain of partially observed multivariate time series by imposing and exploiting additional information encoded in the underlying dynamical model. Specifically, we assume the data-generating process is a structural vector autoregression (SVAR) with latent components.

A novel feature of our approach is that we allow for both "contemporaneous" causal influence and arbitrary latent confounding in the data-generating process. The nearest method to ours is the tsFCI algorithm of Entner and Hoyer (2010), which excludes the possibility of contemporaneous influence. So, our paper begins with an extended characterization of the target data-generating processes and the challenges they present. We then summarize related work on causal structure learning from time series before introducing our

algorithms: SVAR-FCI and SVAR-GFCI. We explore the performance of our methods with simulation experiments and apply SVAR-FCI to some real data from international political economy.

2. Preliminaries

A graph $\mathcal{G}$ is a pair $(\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a set of vertices corresponding to random variables (e.g., $\mathbf{V} = \{X_1, ..., X_k\}$) and $\mathbf{E}$ is a set of edges connecting vertices in $\mathbf{V}$.

Definition 2.1 *Basic graphical terminology. If $X_i \to X_j$ then X_i is called a* parent *of X_j, and X_j is a* child *of X_i. Two variables are* adjacent *if there is some edge between them, and a* path *is a sequence of distinct adjacent vertices containing at least two vertices, e.g., $\langle X_i, X_{i+1}, ..., X_{i+n} \rangle$. A path is a* directed path from X_i to X_n *if for all $m \in \{1, ..., n\}$ the edge $X_{i+m-1} \to X_{i+m}$ occurs. When there is a directed path from X_i to X_j we call X_i an* ancestor *of X_j, and X_j is a* descendent *of X_i. Denote the set of parents of a vertex X in $\mathcal{G}$ by $pa(X, \mathcal{G})$ and the set of adjacencies of X by $adj(X, \mathcal{G})$. In an* acyclic *graph, no vertex is an ancestor (nor a descendent) of itself.*

A DAG is a graph that contains only directed edges ($\to$) and which is acyclic. In a causal DAG, $X_i \to X_j$ if and only if X_i is a direct cause of X_j relative to $\mathbf{V}$. There is a straightforward correspondence between direct causation in causal DAG models and (nonparametric) structural equation models which are employed in many sciences including economics, sociology, biology, and so on. For a recent overview of structural equations and their causal interpretation see Peters et al. (2017). The direct causes of a variable X_i appear in the structural equation for X_i, along with a stochastic error term:

$$X_i = f_i(pa(X_i, \mathcal{G}), \varepsilon_i) \tag{1}$$

$\forall i \in \{1, ..., k\}$, i.e., for all vertices in $\mathbf{V}$. The $\varepsilon_1, ..., \varepsilon_k$ are jointly independent and $\varepsilon_i \perp\!\!\!\perp pa(X_i, \mathcal{G})$. f_i can be any measurable function. Note that acyclic graphs correspond to recursive systems of structural equations, i.e., systems with no causal "feedback." Non-recursive structural equations can be represented with cyclic directed graphs (Spirtes, 1995). The conditional independence relationships implied by a recursive structural equation model can be obtained from the corresponding DAG using the well-known d-separation criterion.

In settings where there may be an unknown number and arrangement of unmeasured confounders (latent common causes), one may rather represent the relations among measured variables by a causal MAG (Maximal Ancestral Graph). A MAG is a mixed graph that may have directed ($\to$) and bidirected ($\leftrightarrow$) edges. More generally, if we include the possibility of selection bias a MAG can also have undirected edges, but we will not consider selection bias here. A MAG represents a DAG (or a set of DAGs sharing common features) after all latent variables have been marginalized out, and it preserves all conditional independence relations among the measured variables which are entailed by the underlying DAG. These can be enumerated via a graphical criterion called m-separation. In a MAG $\mathcal{M}$, a tail mark at X_i (e.g., $X_i \to X_j$) means that X_i is an ancestor of X_j in all DAGs represented by $\mathcal{M}$. An arrowhead at X_i (e.g., $X_i \leftarrow X_j$ or $X_i \leftrightarrow X_j$) means that X_i is not an ancestor of X_j in all DAGs represented by $\mathcal{M}$. A $\leftrightarrow$ edge between two variables

indicates that neither variable is an ancestor of the other. Adjacencies in $\mathcal{M}$ occur when X_i and X_j are not d-separated by any subset of the observed variables in all DAGs represented by $\mathcal{M}$. See Richardson and Spirtes (2002), Ali et al. (2009), and Zhang (2008a) for details on MAGs. A Markov equivalence class of MAGs (i.e., a set of MAGs that imply the same m-seperation facts) is represented by a PAG (Partial Ancestral Graph), which possibly has edges with the additional "circle" edge mark $\circ$ (e.g., $X_i \circ\!\!\rightarrow X_j$). This indicates that in some MAG in the equivalence class there is an arrowhead at X_i and in some other MAG there is a tail at X_i. So, the PAGs we will consider in this paper (again, excluding the possibility of selection variables) can have the following edges: $\rightarrow$, $\circ\!\!\rightarrow$, $\circ\!\!-\!\!\circ$, and $\leftrightarrow$.

2.1 Dynamic DAGs with latent variables

The above correspondence scheme between DAGs and structural equations can be extended to dynamic systems, where vertices represent elements of discrete-time stochastic processes in lieu of cross-sectional observations and the equations are (possibly nonlinear) SVARs. The present focus will be on stochastic processes $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ that are generated by SVARs including arbitrary latent components. A k-dimensional order-p SVAR process may be written:

$$X_{i,t} = f_i(\mathbf{X}_t^{-i}, \mathbf{X}_{t-1}, ..., \mathbf{X}_{t-p}, \varepsilon_{i,t}) \tag{2}$$

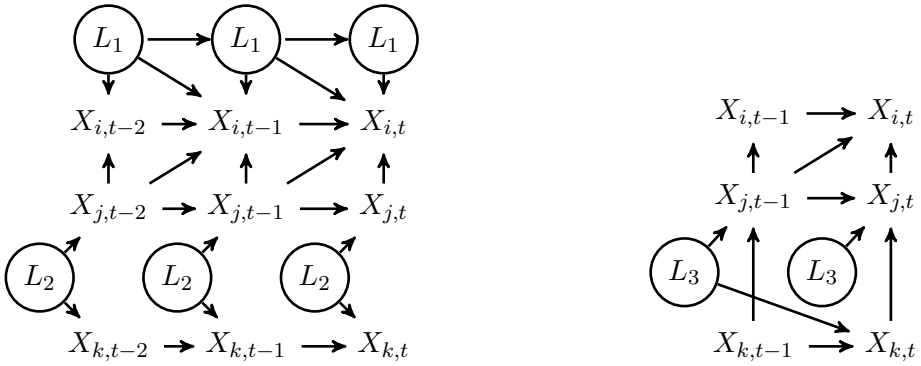
$\forall i \in \{1, ..., k\}, \forall t \in \mathbb{N}$. $\mathbf{X}_t$ is a $k \times 1$ vector of time series variables $(X_{1,t}, ..., X_{k,t})'$, $\mathbf{X}_t^{-i} = \mathbf{X}_t \setminus \{X_{i,t}\}$, and the $\varepsilon_{i,t}$ are both mutually and serially independent (cf. Peters et al., 2013). Just as for the structural equation models introduced earlier, the f_i can be arbitrary measurable functions. In the linear case it is more common to write in matrix notation:

$$\mathbf{\Gamma}_0\mathbf{X}_t = \mathbf{\Gamma}_1\mathbf{X}_{t-1} + ... + \mathbf{\Gamma}_p\mathbf{X}_{t-p} + \boldsymbol{\varepsilon}_t \tag{3}$$

$\forall t \in \mathbb{N}$ where the $\mathbf{\Gamma}_j$ are $k \times k$ matrices of constant coefficients and $\mathbf{\Gamma}_0$ has ones along the diagonal. If the errors are jointly normal, the independent errors assumption implies that $\mathbf{\Sigma} = \mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t']$ is diagonal. Such models are especially common in empirical macroeconomics. We make one important substantive restriction: we exclude the possibility of contemporaneous causal feedback, i.e., the f_i in (2) must be recursive with respect to variables at t and $\mathbf{\Gamma}_0$ can be made lower triangular in (3). The data-generating process is permitted to have any number of latent components; these are sometimes explicitly represented by replacing $\mathbf{X}_t$ in (2) or (3) with $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$, $\mathbf{L}_t$ being a vector of unmeasured time series variables and $\mathbf{X}_t$ being the observed variables (likewise for all other terms). Throughout we assume that the process is stable and thus stationary, i.e., that all moments of the process are time invariant. In the linear case, the stability condition is that all roots of the reverse characteristic polynomial are outside the complex unit circle (Lütkepohl, 2005, p. 16). In the general nonlinear case, stability conditions become much more complicated (Saikkonen, 2001).

Corresponding to such a data-generating process is a dynamic DAG with latent variables, also called a dynamic Bayesian network (DBN).[1] If the function f_i is a nontrivial function of $X_{j,s}$ $(s \leq t)$, then $X_{j,s} \rightarrow X_{i,t}$ in the graph $\mathcal{G}$. Note that the graphs considered here can be

1. We prefer the term "dynamic DAG" to DBN because DBNs are typically specified in two parts: a prior network and a transition network (Friedman et al., 1998). The dynamic DAG is just one graph here, and this is sometimes called the "unrolled DBN."



(a) (b)

Figure 1: a) A dynamic DAG model with latent processes L_1 and L_2. L_1 may be called an "auto-lag confounder," and L_2 may be called a "contemporaneous confounder." b) A dynamic DAG model with latent process L_3. L_3 may be called a "cross-lag confounder."

called *repeating* since $X_{i,t-h} \to X_{j,t}$ if and only if $X_{i,t-h-m} \to X_{j,t-m}$ $\forall h \geq 0, m \in \mathbb{N}$. Also note that these graphs are (semi-)infinite, i.e., $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ where $|\mathbf{V}| = |\mathbb{N} \times \tilde{\mathbf{X}}_t|$. However, in practice we handle only finite segments of these graphs since they are repeating. Two examples are displayed in Figure 1. The graph segments displayed here have the same edges among variables at $t-p$ as among variables at t. This representational choice is not uniform across studies; in some presentations, dynamic DAGs are drawn with "extra" edges at time-slice $t-p$ (i.e., edges that do not appear at slice t) because some variables at $t-p$ are conditionally independent only given some lagged covariates outside of the visible segment. Keeping the infinite repeating graph in mind, we find it more convenient to represent finite segments with the same edges at $t-p$ and t.

Given initial values $(\mathbf{X}_{-p+1}, ..., \mathbf{X}_0)'$, the Markov factorization for a k-dimensional order p stochastic process $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$ may be written:

$$p(\mathbf{X}_t, ..., \mathbf{X}_{t-p}) = \prod_{i \in \{1,...,k\},\ s \in \{t,...,t-p\}} p(X_{i,s} | pa(X_{i,s}, \mathcal{G})) \tag{4}$$

$\forall t \in \mathbb{N}$, where $\mathcal{G}$ is the infinite dynamic DAG.

2.2 Dynamic ancestral graph Markov models

We assume the data-generating process is some SVAR (2) for $\{\tilde{\mathbf{X}}_t\}_{t \in \mathbb{N}}$, $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$ with order p, and that we observe samples from the marginal subprocess $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$. For the dynamic DAG segment $\mathcal{G}$ corresponding to the entire process we can derive a unique dynamic MAG segment $\mathcal{M}$ over just the observed variables. Two examples are illustrated in Figure 2. The algorithms introduced in Section 3 aim to learn the Markov equivalence class of $\mathcal{M}$, a dynamic PAG segment $\mathcal{P}$.

Latent variables in dynamic DAGs may be classified into several types: there may be "contemporaneous confounders" which are common causes of multiple contemporaneous variables (L_2 in Figure 1a); there may be "cross-lag confounders" which are common causes

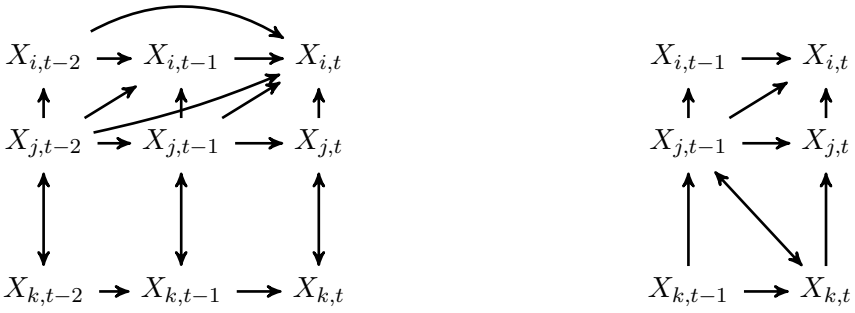


Figure 2: a) The dynamic MAG model implied by Figure 1a. b) The dynamic MAG model implied by Figure 1b.

of variables at different lags, say $X_{k,t}$ and $X_{j,t-1}$ (L_3 in Figure 1b); finally there may be "auto-lag confounders" which are common causes of some variable and it's own past lags, say $X_{i,t-1}$ and $X_{i,t}$ (L_1 in Figure 1a). Latent variables may be combinations of these types; they also may or may not cause themselves (e.g., $L_{1,t-1} \to L_{1,t}$) and cause each other. Each type of confounding poses difficulties for analysis, but auto-lag confounders can be particularly problematic since they induce apparently "infinite-lag" associations in the marginal: $X_{i,t}$ (in Figure 1a) will not be conditionally independent of its distant past even conditional on any number of lags, which makes the marginal model appear to be of infinite order. This is important because in practice a user of some causal structure learning method must choose a finite number of lags to include their analysis, which may not reflect the true Markov order of the fully observed process. For example, having chosen $p = 2$ the MAG in Figure 2a includes the edge $X_{i,t-2} \to X_{i,t}$ though the underlying dynamic DAG has order 1. Taking care not to mistake the apparent "order" of the MAG for the order of the underlying DAG, we note the MAG when properly interpreted preserves relevant causal and statistical information: $X_{i,t-2}$ is indeed a causal ancestor of $X_{i,t}$ in the underlying DAG; $X_{i,t-2}$ and $X_{i,t}$ are not d-separated by any subset of the observed variables in the underlying DAG; and applying the distinction between "visible" and "invisible" edges in a MAG produces the conclusion that a possible latent confounder may be affecting $X_{i,t-2}$ and $X_{i,t}$, which is true (Zhang, 2008a). We discuss a data-driven way of choosing the number of lags to include for structure learning in the Appendix.

In contrast to previous work on time series (discussed below), we allow for arbitrary unmeasured structure. The algorithms introduced in this paper make no attempt to discover the number, arrangement, or features of the unmeasured processes – rather our methods learn as much as possible (from independence constraints) about the causal relations among what is observed. Just as with FCI and GFCI, we aim to exploit patterns of conditional independence and dependence in the marginal distribution to rule out confounding among some of the variables, and thus render some (if not all) of the causal effect parameters estimable. If there is only one or a few causal relationships of particular interest and confounding cannot be ruled out on the basis of the data, this may indicate to a researcher

that they must either make stronger assumptions to procure unbiased estimates of the desired quantity, or expand their set of measured variables.

2.3 Contemporaneous causal relations

The data-generating processes considered here are discrete-time autoregressive models with contemporaneous causal influences. Popular approaches to modeling such data, e.g., so-called Granger-causality methods or VAR estimation, make no reference to contemporaneous relations or sometimes explain-away contemporaneous statistical dependencies by non-causal means. It is worth considering these issues in more detail before proceeding, since contemporaneous relations may in fact be interpreted in multiple ways.

In SVARs, contemporaneous causal relations (such as summarized in the matrix $\mathbf{\Gamma}_0$ in the linear case) represent the possibility that interventions (e.g., "shocks" typically studied in macroeconometrics) on one variable can have consequences for another variable "within the same period," i.e., that causal influences can propogate more quickly than the frequency of available measurements. Note that this viewpoint contrasts with Granger's (1988, p. 205-6), who prefers to explain all contemporaneous statistical dependence as an artifact of unmeasured confounding: on his view $X_{i,t}$ does not cause $X_{j,t}$ contemporaneously, but rather there is some unmeasured common cause of both variables $L_{t-\delta}$ ($\delta > 0$) which is determined at an earlier time.[2] The two cases have distinguishable consequences: on a missing variables interpretation, interventions on $X_{i,t}$ would have no effect on $X_{j,t}$ (but may have effect on $X_{j,t'}$ for some later t'), whereas if the data-generating process for the complete system (i.e., including latent processes) is allowed to have contemporaneous dependencies, then interventions on $X_{i,t}$ may have consequences for $X_{j,t}$. The algorithms introduced below can distinguish between such hypotheses, since dynamic MAGs are permitted to have either bidirected or directed edges between contemporaneous variables.

Causal influence "within the same period" or different speeds of influence propogation can be explicated by considering subdivisions of the observation period, as well as more substantially delayed relations. A discrete time window may be divided into n equal intervals, with $\Delta\theta = 1/n$ characterizing the underlying speed of causal propogation within the window. That is, one may consider the underlying multi-scale model (shifting notation from $X_{i,t}$ to $X_i(t)$ for readability):

$$\mathbf{X}(t + k\Delta\theta) \quad = \quad \mathbf{B}_0\mathbf{X}(t + (k-1)\Delta\theta) + \sum_{\tau=\tau_1}^{\tau_p} \mathbf{\Gamma}_\tau \mathbf{X}(t-\tau) + \boldsymbol{\varepsilon}(t + k\Delta\theta) \tag{5}$$

$\forall t \in \mathbb{N}$, $k = 1, ..., n$. It is assumed that the delay points $\{\tau_1, ..., \tau_p\}$ are multiples of the step size. Here we have no contemporaneous relations.

Plausibly, the measurement frequency for the observed time series is much lower than the incremental step size $\Delta\theta$. On one interpretation connecting the "measurement timescale" to

2. In the same paper, Granger also claims that observed contemporaneous causal relations can be "explained by either temporal aggregation or missing causal variables," but he does not discuss temporal aggregation further and focuses on the missing variables explanation (p. 206). In Granger's (1969), he suggests that "real" (as contrasted with "spurious") contemporaneous causality can be attributed to the relative frequency of measurements and speed of information flows (p. 427 and 430). In earlier work (1963, p. 41) Granger explicates this relative frequency interpretation with a kind of undersampling scheme, discussed below.

the "fast timescale" of the multi-scale model, the SVAR which is the focus of this work arises naturally from (5). Following the strategy in Fisher (1970), consider observations which are local time averages (see also Gong et al., 2017). Fisher concerns himself with the simultaneous equation models popular among mid-20th century econometricians, but one may extend his strategy to the case of linear SVARs. For convenience of exposition, say that the measurements are collected monthly but the underlying causal influences propogate on the scale of a second. So, the observation interval (a month) may be divided into n equal subintervals each of length $\Delta\theta$ (approx. one second). Fisher posits that the measured variables are actually time averages over this observation period. In other words, we measure variables like GDP and consumption as averages over a month, $X_{i,t} := \bar{X}_i(t) = \frac{1}{n}\sum_{k=1}^{n} X_i(t + k\Delta\theta)$. In that case, the multi-scale model (5) can be locally averaged to yield a SVAR relating the average variables in the limit as $n \to \infty$ or $\Delta\theta \to 0$. We begin by aligning things so $\tau_1 = 1, ..., \tau_p = p$. Then, following Fisher and using the above definition of time average,

$$\begin{aligned}
\bar{\mathbf{X}}(t) &= \frac{1}{n}\sum_{k=1}^{n}\{\mathbf{B}_0\mathbf{X}(t + (k-1)\Delta\theta) + \sum_{j=1}^{p}\boldsymbol{\Gamma}_j\bar{\mathbf{X}}(t-j) + \boldsymbol{\varepsilon}(t + k\Delta\theta)\} \\
&= \mathbf{B}_0\frac{1}{n}\sum_{k=1}^{n}\mathbf{X}(t + (k-1)\Delta\theta) + \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{p}\boldsymbol{\Gamma}_j\bar{\mathbf{X}}(t-j) + \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{\varepsilon}(t + k\Delta\theta) \\
&= \mathbf{B}_0\frac{1}{n}\sum_{k=1}^{n}\mathbf{X}(t + k\Delta\theta) + \sum_{j=1}^{p}\boldsymbol{\Gamma}_j\bar{\mathbf{X}}(t-j) + \bar{\boldsymbol{\varepsilon}}(t) + \mathbf{B}_0\frac{1}{n}\mathbf{X}(t) - \mathbf{B}_0\frac{1}{n}\mathbf{X}(t + n\Delta\theta) \\
&= \mathbf{B}_0\bar{\mathbf{X}}(t) + \sum_{j=1}^{p}\boldsymbol{\Gamma}_j\bar{\mathbf{X}}(t-j) + \bar{\boldsymbol{\varepsilon}}(t) + \mathbf{B}_0\frac{1}{n}(\mathbf{X}(t) - \mathbf{X}(t + n\Delta\theta))
\end{aligned}$$

The last equation yields the linear SVAR (3) with $\mathbf{B}_0 = (\mathbf{I} - \boldsymbol{\Gamma}_0)$ in the limit as $n \to \infty$ or $\Delta\theta \to 0$ if and only if

$$\lim_{n\to\infty}\frac{1}{n}(\mathbf{X}(t) - \mathbf{X}(t + n\Delta\theta)) = \lim_{\Delta\theta\to 0}\Delta\theta(\mathbf{X}(t) - \mathbf{X}(t + n\Delta\theta)) = 0. \tag{6}$$

This is Fisher's equation (3.4) in (1970). From here Fisher goes on to show that to satisfy this condition, it is necessary and sufficient that $|\lambda_i| \leq 1$, where λ_i $(i = 1, ..., k)$ are the characteristic values of $\mathbf{B}_0$. $|\cdot|$ denotes the modulus. In nonlinear systems, the relevant functions have to be restricted analagously (e.g., their derivatives must be bounded in absolute value).

So, the linear SVAR model (3) may arise from a local average approximation to an underlying multi-scale model, in the limit as $n \to \infty$. Note that Fisher (1970) restricts the error variables to be constant over the observation interval, whereas here the error variables in (3) are treated as time averages on par with all the other variables in the model. Of course, not all variables are local averages, and not in all settings is the $n \to \infty$ assumption reasonable.

In some settings, it may be more appropriate to think of observed data as generated by an undersampling (a.k.a. subsampling) scheme, i.e. that the data consists of every mth observation of (5). This may be the case, for example, in neurological studies where data-generating processes are fast and measurement technology is precise but much slower. In

a sense, it is possible to interpret contemparaneous correlation in a VAR model as an artifact of undersampling: undersamping a VAR process with no contemporaneous relations (i.e., an SVAR with $\mathbf{\Gamma}_0 = \mathbf{I}$) would indeed produce contempraneous correlation among the residuals, and one may apply traditional econometric techniques to render those residuals uncorrelated, e.g., premultiplying by some orthogonalizing matrix (Lütkepohl, 2005). However, it is important to note that the parameters in such a model – regardless of how the orthogonalizing matrix is chosen – are not the underlying structural parameters in the "fully sampled" representation. Unlike the local averaging case, undersampling does not preserve causal relationships, and one cannot in general straightforwardly infer the causal relationships in the underlying VAR by estimating an SVAR model for the undersampled data. See Gong et al. (2015) for detailed discussion of this point.

So, we do not interpret the contemporaneous relations in the SVAR (2) or (3) data-generating process as an artifact of undersampling, though the approach outlined in this paper accommodates undersampling in a different way. Namely, the unrecorded observations in an undersampled process may be treated as latent variables, and searching for the causal structure of such a system is just equivalent to searching for the induced marginal model of an SVAR with latent variables (a dynamic MAG) in the special case where the latent variables are the excluded observations of the measured variables. It may be helpful to consider a simple example with just two processes. Consider the model (shifting notation back to $X_{i,t}$ form):

$$\begin{aligned} X_{i,t} &= \gamma_1 X_{i,t-1} + \varepsilon_{i,t} \\ X_{j,t} &= \gamma_2 X_{i,t-1} + \gamma_3 X_{j,t-1} + \varepsilon_{j,t} \end{aligned} \tag{7}$$

Undersampling this model by a factor of two ($m = 2$) means that we observe every second value. The intervening variables are latent, so the model is equivalent to:

$$\begin{aligned} L_{i,t-1} &= \tilde{\gamma}_1 X_{i,t-2} + \tilde{\varepsilon}_{L_i,t-1} \\ L_{j,t-1} &= \tilde{\gamma}_2 X_{i,t-2} + \tilde{\gamma}_3 X_{j,t-2} + \tilde{\varepsilon}_{L_j,t-1} \\ X_{i,t} &= \tilde{\gamma}_4 L_{i,t-1} + \tilde{\varepsilon}_{i,t} \\ X_{j,t} &= \tilde{\gamma}_5 L_{i,t-1} + \tilde{\gamma}_6 L_{j,t-1} + \tilde{\varepsilon}_{j,t} \end{aligned} \tag{8}$$

This corresponds to a dynamic DAG with latent variables. In matrix notation (8) is written:

$$\tilde{\mathbf{\Gamma}}_0 \tilde{\mathbf{X}}_t = \tilde{\mathbf{\Gamma}}_2 \tilde{\mathbf{X}}_{t-2} + \tilde{\boldsymbol{\varepsilon}}_t \tag{9}$$

where $\tilde{\mathbf{X}}_t = (L_{i,t-1}, L_{j,t-1}, X_{i,t}, X_{j,t})'$. So we have a special case of the SVAR model with latent processes. The methods described in the next section accommodate undersampling by learning the structure of the marginal model, in contrast to approaches which learn the parameters or structure of the underlying "fully sampled" process from undersampled data (e.g., Plis et al., 2015; Gong et al., 2015; Hyttinen et al., 2017).

So, the various ways of understanding contemporaneous causal connections – causal influences which propogate faster than the measurement time-scale, whether the measurements are local averages or undersampled observations, with or without latent processes – are accommodated within the general model class we consider: the class of SVARs with latent components.

2.4 Related work

Swanson and Granger (1997) search for an SVAR with some restrictions on the possible contemporaneous structure. After estimating a reduced form VAR model in the usual way (multivariate regression of each variable on all earlier variables), they find correlated errors and Swanson and Granger propose to explain these by contemporaneous causal connections organized in a "causal chain." A causal chain is a recursive total causal ordering, like $X_{1,t} \to X_{2,t} \to X_{3,t}$. Variables which are (contemporaneous) common causes or common effects of multiple variables are excluded from the space of possibilities. In summary, Swanson and Granger employ a two-step strategy – regression estimates of the reduced form VAR, followed by identification of the contemporaneous relations – to recover an SVAR like (3) with some restrictions on $\mathbf{\Gamma}_0$. Generalizations of this two-step procedure are pursued later in the econometric literature. Bessler and Lee (2002) and Demiralp and Hoover (2003) opt for the more unrestricted domain of recursive causal orderings for $\mathbf{\Gamma}_0$; they use the PC algorithm which allows for any recursive (acyclic) structure. Hyvärinen et al. (2010) and Moneta et al. (2011) (see also works cited therein) pursue essentially the same strategy except they use ICA-LiNGaM to search for contemporaneous orientations. ICA-LiNGaM allows for any recursive ordering and assumes that the error distributions are non-Gaussian. All of these studies assume linearity. More seriously, all of these methods exclude latent variables, although one could in principle replace the PC algorithm and ICA-LiNGAM with something like FCI or LV-LiNGaM to allow for latent confounding among contemporaneous variables. This approach would still leave out the possibility of hidden common causes among the lagged variables, namely what we call cross-lag and auto-lag confounding in Figure 1. Chu and Glymour (2008) do something along these lines by using FCI for contemporaneous connections but they use additive regression techniques for the non-contemporaneous connections, so they allow for only contemporaneous confounding but no cross-lag or auto-lag confounding. Their approach is different from the others in that they allow for additive but nonlinear relations among the variables, and allow for unmeasured confounders even if only in a limited way.

Eichler (2012) models multivariate time series with mixed graphical models which are very similar to the ancestral models proposed here. However, Eichler's representation suppresses information about the dynamics of the process; his framework represents causal connections between entire time series processes but cannot distinguish between (for example) the case where a variable causes another at one time lag ($X_{i,t-1}$ causes $X_{i,t}$) and the case where a variable causes another at only two time lags ($X_{i,t-2}$ causes $X_{i,t}$). Eichler (2012, p. 10-11) points out that representing the full dynamics can be computationally complex, so he prefers the simpler and more tractable representation. On the other hand, representing the dynamics (the lag structure) of a causal system can be critical for forecasting the outcomes of policies as in impulse response analysis or assessing dynamic treatment strategies.

The approach we present here follows Entner and Hoyer (2010) in searching for (equivalence classes of) ancestral graphs over the full dynamic structure. A key difference between SVAR-FCI as presented below and their tsFCI algorithm is that we allow for contemporaneous causal relationships whereas Entner and Hoyer do not; we elaborate in the next section. In addition, we introduce a (hybrid) score-based approach which Entner and Hoyer

do not explore. Gao and Tian (2010) present a likelihood-based procedure for selecting dynamic ancestral graph models but their approach suffers from several shortcomings. First, by performing model selection using maximum likelihood with no complexity penalty, their procedure will select overly complex models even in the limit of infinite data. That is, their procedure is not consistent. Second, they perform structure learning directly on MAGs rather than PAGs, thus ignoring problems of Markov equivalence; their procedure selects the model with maximum likelihood despite the fact that several alternative models will be likelihood-equivalent given their assumptions. Third, they perform a brute force enumeration of all dynamic MAGs over the measured variables in their search-and-score procedure, which is computationally infeasible for more than a few variables.

3. Learning ancestral models

We modify the FCI algorithm to search for an equivalence class of dynamic MAGs. Assuming that the data-generating process is a (possibly nonlinear) SVAR with latent variables, or equivalently a dynamic DAG with latent components, the SVAR-FCI procedure makes the following modifications to FCI:

1) SVAR-FCI respects the time order of the variables by restricting possible conditioning sets to variables in the "present" or "past" time slices and prohibiting orientations backwards in time.

2) SVAR-FCI enforces the repeating structure of the underlying dynamic DAG in both determining adjacencies and orientations.

Following Entner and Hoyer (2010), we introduce the following definition:

Definition 3.1 *Let the pair of vertices* $(X_{i,t}, X_{j,s})$ *be called* homologous *to pair* $(X_{m,a}, X_{n,b})$ *if* $m = i$, $n = j$, *and* $t - s = a - b$. $hom(X_{i,t}, X_{j,s}, \mathcal{G})$ *denotes the set of vertex pairs homologous to* $(X_{i,t}, X_{j,s})$ *in graph* $\mathcal{G}$.

For example, in Figure 1a the pairs $(X_{i,t-1}, X_{j,t-1})$ and $(X_{i,t-2}, X_{j,t-2})$ are homologous to $(X_{i,t}, X_{j,t})$; $(X_{k,t-2}, X_{j,t-1})$ is homologus to $(X_{k,t-1}, X_{j,t})$. SVAR-FCI enforces the repeating structure of the assumed data-generating process by removing the edge between all $(X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s})$ whenever an edge is removed from $(X_{i,t}, X_{j,s})$, and orienting the edge between all $(X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s})$ as the edge between $(X_{i,t}, X_{j,s})$ is oriented. We also need the following definitions:

Definition 3.2 *Given a path* π *in a graph* $\mathcal{G}$, *a non-endpoint vertex* X_j *on* π *is called a* collider *if the two edges incident to* X_j *are both into* X_j, *i.e., have arrowheads at* X_j $(*\rightarrow X_j \leftarrow *)$. *(Note the* $*$ *mark is used to represent any possible endpoint.) A* v-structure *is a triple* $\langle X_i, X_j, X_k \rangle$ *such that* $X_i *\rightarrow X_j \leftarrow * X_k$ *and* X_i *and* X_k *are not adjacent.*

Definition 3.3 *Let* $X \in pds(X_i, X_j, \mathcal{G})$ *if and only if* $X \neq X_i$, $X \neq X_j$, *and there is a path* π *between* X_i *and* X *in* $\mathcal{G}$ *such that for every subpath* $\langle X_m, X_l, X_h \rangle$ *of* π *either* X_l *is a collider on the subpath in* $\mathcal{G}$ *or* $\langle X_m, X_l, X_h \rangle$ *is a triangle in* $\mathcal{G}$. *A* triangle *is a triple* $\langle X_m, X_l, X_h \rangle$ *where each pair of vertices is adjacent.*

Also, define $adj_t(X_{i,t}, \mathcal{G}) = \{X_{j,s} : X_{j,s} \in adj(X_{i,t}, \mathcal{G}), s \leq t\}$ and $pds_t(X_{i,t}, X_{k,u}, \mathcal{G}) = \{X_{j,s} : X_{j,s} \in pds(X_{i,t}, X_{k,u}, \mathcal{G}), s \leq \max(t, u)\}$. Psuedocode for SVAR-FCI is presented in Algorithm 3.1 below. (We include pseudocode for FCI in the Appendix for comparison.)

Note that our procedure differs from the tsFCI algorithm presented by Entner and Hoyer (2010) in two ways: first, we do not restrict contemporaneous causal connections, except that we disallow cycles; second, we remove adjacencies and propogate orientations for homologous edges even if the separating set for $(X_{i,t}, X_{j,s})$ is outside the visible window (so, as a consequence, the structure at time-slice $t-p$ is the same as at t). The latter change may appear minor, but is helpful for interpreting the output as the marginal ancestral graph of an infinite dynamic DAG, and it significantly increases the number of unambiguous orientations, since orientation decisions at earlier time-slices can help orient later edges according to the rules in Zhang (2008b). The first difference is also important: we allow for data-generating processes with nontrivial contemporaneous causation, whereas Entner and Hoyer attribute all observed contempraneous dependence to latent confounding. In this sense, SVAR-FCI as presented here places fewer restrictions on the possible underlying data-generating processes.

SVAR-FCI is consistent under the same conditions as the FCI algorithm: the Markov condition and faithfulness (Spirtes et al., 2000; Zhang, 2008b), though in this case w.r.t. an underlying dynamic DAG. We require a consistent test of conditional independence for time series data, i.e., a test which makes correct conditional independence judgements in the limit as sample size $T \to \infty$.

Proposition 3.1 *Assume the stationary stochastic process $\{\tilde{\mathbf{X}}_t\}_{t\in\mathbb{N}}$, where $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$, is Markov and faithful to a dynamic DAG $\mathcal{G}$. Let $\mathcal{M}$ be the MAG implied by $\mathcal{G}$ over $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ and PAG $\mathcal{P}$ the equivalence class of $\mathcal{M}$. Given T observations of the marginal subprocess $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ and a consistent test of conditional independence, the SVAR-FCI algorithm is a (pointwise) consistent estimator of $\mathcal{P}$.*

Proof This follows straightfowardly from the consistency of the FCI algorithm. The task is equivalent to estimating the PAG $\mathcal{P}$ over vertex set $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$, with the added assumptions that the underlying DAG is repeating and that there are no directed edges from later vertices to earlier vertices. The SVAR-FCI algorithm just enforces these restrictions at every step of the FCI algorithm. ■

SVAR-FCI may be considered a nonparametric method if used in conjunction with a nonparametric test consistent for non-i.i.d. data. Some nonparametric conditional independence tests for multivariate time series are discussed and evaluated in Moneta et al. (2011). Unfortunately, the methods described therein require multivariate kernel density estimation and a complicated bootstrap procedure to calculate critical values, so these tests are so far infeasible for settings with many variables. Scalable nonparametric independence tests are a subject of ongoing research. For linear processes with Gaussian errors, the Fisher Z test of vanishing partial correlation is a popular and fast choice. We discuss a data-driven selection procedure for the tuning parameter α, the nominal test size, in the Appendix. A similar procedure can be used for selecting the order p, which we also discuss in the Appendix.

Algorithm 3.1: SVAR-FCI(Test, α)

Input: Data on variables $\mathbf{X}_t, ..., \mathbf{X}_{t-p} = \{X_{1,t}, ..., X_{k,t}, ..., X_{1,t-p}, ..., X_{k,t-p}\}$
Output: Dynamic PAG segment $\mathcal{P}$
1. Form the complete graph $\mathcal{P}$ on vertex set $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ with $\circ\!\!-\!\!\circ$ edges.
2. $n \leftarrow 0$
3. **repeat**
4. **for all** pairs of adjacent vertices $(X_{i,t}, X_{j,s})$ s.t. $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| \geq n$ and subsets $\mathbf{S} \subset adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}$ s.t. $|\mathbf{S}| = n$
5. **if** $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)
then { Delete edge $X_{i,t} \circ\!\!-\!\!\circ X_{j,s}$ from $\mathcal{P}$. Delete edge $X_{m,a} \circ\!\!-\!\!\circ X_{n,b}$ $\forall (X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s}, \mathcal{P})$. Let $sepset(X_{i,t}, X_{j,s}) = sepset(X_{i,t}, X_{j,s}) = \mathbf{S}$.
6. **end**
7. $n \leftarrow n + 1$
8. **until** for each pair of adjacent vertices $(X_{i,t}, X_{j,s})$, $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| < n$.
9. **for all** adjacent vertices $(X_{i,t}, X_{j,s})$ orient $X_{i,t} \ast\!\!\rightarrow X_{j,s}$ iff $s > t$.
10. **for all** triples $(X_{i,t}, X_{k,r}, X_{j,s})$ s.t. $X_{i,t} \in adj_t(X_{k,r}, \mathcal{P})$ and $X_{j,s} \in adj_t(X_{k,r}, \mathcal{P})$ but $X_{i,t} \notin adj_t(X_{j,s}, \mathcal{P})$, orient $X_{i,t} \ast\!\!\rightarrow X_{k,r} \leftarrow\!\!\ast X_{j,s}$ iff $X_{k,r} \notin sepset(X_{i,t}, X_{j,s})$; then also orient $X_{m,a} \ast\!\!\rightarrow X_{o,c} \leftarrow\!\!\ast X_{n,b}$ $\forall (X_{m,a}, X_{o,c}) \in hom(X_{i,t}, X_{k,r}, \mathcal{P})$ and $\forall (X_{n,b}, X_{o,c}) \in hom(X_{j,t}, X_{k,r}, \mathcal{P})$
11. **for all** pairs $(X_{i,t}, X_{j,s})$ adjacent in $\mathcal{P}$ **if** $\exists \mathbf{S}$ s.t. $\mathbf{S} \in pds_t(X_{i,t}, X_{j,s}, \mathcal{P})$ or $\mathbf{S} \in pds_s(X_{j,s}, X_{i,t}, \mathcal{P})$ and $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)
then { Delete edge $X_{i,t} \circ\!\!-\!\!\circ X_{j,s}$ from $\mathcal{P}$. Delete edge $X_{m,a} \circ\!\!-\!\!\circ X_{n,b}$ $\forall (X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s}, \mathcal{P})$. Let $sepset(X_{i,t}, X_{j,s}) = sepset(X_{i,t}, X_{j,s}) = \mathbf{S}$.
12. Reorient all edges as $\circ\!\!-\!\!\circ$ and **repeat** steps 9 and 10.
13. Exhaustively apply orientation rules (R1-R10) in Zhang (2008b) to orient remaining endpoints, orienting all homologous pairs similarly.
14. **return** $\mathcal{P}$.

In i.i.d. settings, simulations have shown that the hybrid score-based method GFCI is typically more accurate in finite samples than its constraint-based counterpart, FCI (Ogarrio et al., 2016). So, one may hope to improve on the accuracy of SVAR-FCI with similar modifications to GFCI, using a greedy initial adjacency search that respects time order and enforces the assumed repeating structure. Pseudocode for SVAR-GFCI is presented in Algorithm 3.2. As a subroutine, we reference SVAR-GES, which modifies GES analagously: SVAR-GES prohibits adding edges inconsistent with the time order of the variables, it adds/removes homologous edges every time an edge is added/removed, and orients homologous edges similarly whenever an edge is oriented. The procedure requires a score that is decomposible, score-equivalent, and consistent (Chickering, 2002). Let $S(\mathcal{G}, \mathbf{D})$ denote the score of model $\mathcal{G}$ with data $\mathbf{D}$. A score is decomposible if $S(\mathcal{G}, \mathbf{D}) = \sum_{X_i \in \mathbf{V}} S(X_i | pa(X_i, \mathcal{G}), \mathbf{D})$, i.e., it can be decomposed into a sum of "local" contributions

from each variable given its parents. A score is score-equivalent if $S(\mathcal{G}, \mathbf{D}) = S(\mathcal{G}', \mathbf{D})$ when $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent. A score is consistent when the true model gets the highest score as sample size approaches infinity. We provide pseudocode for GFCI and important subroutines of GES in the Appendix, to aid in comparison.

Algorithm 3.2: SVAR-GFCI(Score, Test, α)

Input: Data on variables $\mathbf{X}_t, ..., \mathbf{X}_{t-p} = \{X_{1,t}, ..., X_{k,t}, ..., X_{1,t-p}, ..., X_{k,t-p}\}$
Output: Dynamic PAG segment $\mathcal{P}$
1. $\mathcal{G} \leftarrow$ SVAR-GES(Score)
2. Form the graph $\mathcal{P}$ on vertex set $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ with adjacencies in $\mathcal{G}$ and $\circ\!\!-\!\!\circ$ edges.
3. $n \leftarrow 0$
4. **repeat**
5. **for all** pairs of adjacent vertices $(X_{i,t}, X_{j,s})$ s.t. $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| \geq n$ and subsets $\mathbf{S} \subset adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}$ s.t. $|\mathbf{S}| = n$
6. **if** $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)
then $\begin{cases} \text{Delete edge } X_{i,t} \circ\!\!-\!\!\circ X_{j,s} \text{ from } \mathcal{P}. \\ \text{Delete edge } X_{m,a} \circ\!\!-\!\!\circ X_{n,b} \; \forall (X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s}, \mathcal{P}). \\ \text{Let } sepset(X_{i,t}, X_{j,s}) = sepset(X_{i,t}, X_{j,s}) = \mathbf{S}. \end{cases}$
7. **end**
8. $n \leftarrow n + 1$
9. **until** for each pair of adjacent vertices $(X_{i,t}, X_{j,s})$, $|adj_t(X_{i,t}, \mathcal{P}) \setminus \{X_{j,s}\}| < n$.
10. **for all** adjacent vertices $(X_{i,t}, X_{j,s})$ orient $X_{i,t} \ast\!\!\rightarrow X_{j,s}$ iff $s > t$.
11. **for all** triples $(X_{i,t}, X_{k,r}, X_{j,s})$ s.t. $X_{i,t} \in adj_t(X_{k,r}, \mathcal{P})$ and $X_{j,s} \in adj_t(X_{k,r}, \mathcal{P})$ but $X_{i,t} \notin adj_t(X_{j,s}, \mathcal{P})$, orient $X_{i,t} \ast\!\!\rightarrow X_{k,r} \leftarrow\!\!\ast X_{j,s}$ iff $(X_{i,t}, X_{k,r}, X_{j,s})$ is a v-structure in $\mathcal{G}$, or it is a triangle in $\mathcal{G}$ and $X_{k,r} \notin sepset(X_{i,t}, X_{j,s})$; then also orient $X_{m,a} \ast\!\!\rightarrow X_{o,c} \leftarrow\!\!\ast X_{n,b}$ $\forall (X_{m,a}, X_{o,c}) \in hom(X_{i,t}, X_{k,r}, \mathcal{P})$ and $\forall (X_{n,b}, X_{o,c}) \in hom(X_{j,t}, X_{k,r}, \mathcal{P})$
12. **for all** pairs $(X_{i,t}, X_{j,s})$ adjacent in $\mathcal{P}$ **if** $\exists \mathbf{S}$ s.t. $\mathbf{S} \in pds_t(X_{i,t}, X_{j,s}, \mathcal{P})$ or $\mathbf{S} \in pds_s(X_{j,s}, X_{i,t}, \mathcal{P})$ and $X_{i,t} \perp\!\!\!\perp X_{j,s} | \mathbf{S}$ according to (Test, α)
then $\begin{cases} \text{Delete edge } X_{i,t} \circ\!\!-\!\!\circ X_{j,s} \text{ from } \mathcal{P}. \\ \text{Delete edge } X_{m,a} \circ\!\!-\!\!\circ X_{n,b} \; \forall (X_{m,a}, X_{n,b}) \in hom(X_{i,t}, X_{j,s}, \mathcal{P}). \\ \text{Let } sepset(X_{i,t}, X_{j,s}) = sepset(X_{i,t}, X_{j,s}) = \mathbf{S}. \end{cases}$
13. Reorient all edges as $\circ\!\!-\!\!\circ$ and **repeat** steps 10 and 11.
14. Exhaustively apply orientation rules (R1-R10) in Zhang (2008b) to orient remaining endpoints, orienting all homologous pairs similarly.
15. **return** $\mathcal{P}$.

Proposition 3.2 *Assume the stationary stochastic process $\{\tilde{\mathbf{X}}_t\}_{t \in \mathbb{N}}$, where $\tilde{\mathbf{X}}_t = (\mathbf{L}_t', \mathbf{X}_t')'$, is Markov and faithful to a dynamic DAG $\mathcal{G}$. Let $\mathcal{M}$ be the MAG implied by $\mathcal{G}$ over $\mathbf{X}_t, ..., \mathbf{X}_{t-p}$ and PAG $\mathcal{P}$ the equivalence class of $\mathcal{M}$. Given T observations of the marginal subprocess $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$, a decomposible, score-equivalent, and consistent score, and a consistent*

test of conditional independence, the SVAR-GFCI algorithm is a (pointwise) consistent estimator of $\mathcal{P}$.

Proof This follows straightforwardly from the consistency of the GFCI algorithm, just as for the previous proposition. ■

We restrict ourselves to the BIC score for multivariate normal distributions here, so SVAR-GFCI is limited to linear SVAR processes with Gaussian errors. However, more general model scores are a subject of current research.

4. Simulation experiments

To investigate the finite-sample performance of SVAR-FCI and SVAR-GFCI, we simulate observations from dynamic DAG models with latent variables and use implementations of our algorithms to recover the corresponding PAGs. These simulations are carried out using the `algcomparison` package for performance evaluation, part of the `TETRAD` software (Ramsey and Malinsky, 2016).[3] In our study, we randomly generate order-1 dynamic DAG models with 10 measured processes either two or four latent confounding processes (corresponding to "moderately confounded" and "highly confounded" settings). The average degrees of the underlying DAGs were 3.75 and 4.36 respectively. The DAGs are parameterized as linear SVARs with Gaussian errors, with all coefficients uniformly selected from the range $\pm[0.30, 0.70]$, and variances uniformly in $\pm[1.0, 3.0]$.

We evaluate SVAR-FCI and SVAR-GFCI by considering "accuracy" as a classification problem, examining precision and recall as a function of sample size. Precision is defined as the ratio of true positives to the total number of positives: TP / (TP + FP). Recall is defined as the ratio of true positives to the total number of true instances: TP / (TP + FN). We can ask about precision and recall for classification of adjacencies or orientations. For example, a true positive adjacency between $X_{i,t}$ and $X_{j,s}$ occurs when the estimated PAG from our structure learning method classifies $X_{i,t}$ and $X_{j,s}$ as adjacent, and $X_{i,t}$ and $X_{j,s}$ are adjacent in the true underlying PAG. Orientation classification consist of two components, tails and arrowheads: in our case, we look at arrowhead precision and recall. The results, averaged over 200 trials, can be found in Figure 3. (Results for tails are qualitatively similar.) In this case, we use the Fisher Z test for conditional independence judgements with $\alpha = 0.01$ in SVAR-FCI and $\alpha = 0.05$ in SVAR-GFCI, and the Gaussian model BIC score in SVAR-GFCI.[4] Note that since there does not exist any competing method which allows for both contemporaneous causation and arbitrary confounding (as detailed in Section 2.4), our simulation experiments serve primarily to illustrate the behavior of our algorithms as a function of sample size and confounding.

3. The implementation used in these simulations has been found to be not complete, thus the results should be viewed as preliminary.
4. We use different values for the tuning parameter α because SVAR-GFCI performs significantly fewer conditional independence tests than SVAR-FCI. The values $\alpha = 0.01$ and $\alpha = 0.05$ were chosen because these were found to have (roughly) the best accuracies for the two algorithms. Formally, to approximate a conditional independence oracle α should decrease as a function of sample size. We display results with a fixed α here to illustrate the consequences of choosing fixed α, as is commonly done in practice.

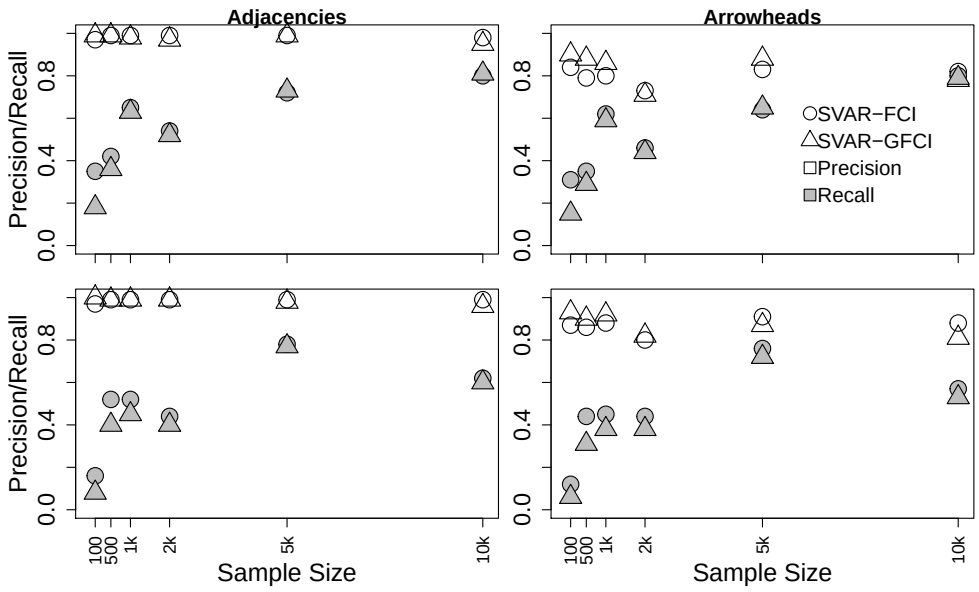


Figure 3: Average precision and recall results in the "moderately confounded" (top row) and "highly confounded" (bottom row) settings, i.e., 10 measured processes and two or four latent processes.

The simulation results paint a familiar picture from constraint-based and score-based structure learning studies in the i.i.d. domain. Both methods have very good adjacency precision, even at small sample sizes. Adjacency recall is generally lower, especially at small sample sizes. The general problem – also apparent in simulation studies with i.i.d. data – is that the estimated graphs are overly sparse, not that they incorrectly include too many edges (Ogarrio et al., 2016). Orientation precision and recall trends are less uniform. Missing adjacencies can lead to wildly different orientations, since these may create incorrect unshielded colliders or other graphical substructures which drive subsequent orientations in the orientation-phase of the algorithms. However, results do indicate that in large samples where the adjacencies are recovered more accurately, arrowhead orientations are not unreasonable. Recall somewhat declines with more latent confounding in the data-generating process, as expected. We also note that neither SVAR-FCI nor SVAR-GFCI seems to be uniformly more accurate with these settings.

5. Application to political economy data

To illustrate how such methods may be applied to real data, we consider date from international political economy on capital taxation rates (Garrett and Mitchell, 2001; Beck and Katz, 2011). The data consists of annual observations from 1967 to 1992 of various macroeconomic and political variables on 16 OECD countries. The total sample size is 330. This is an example of so-called time-series cross-sectional data (TSCS) which is typical of

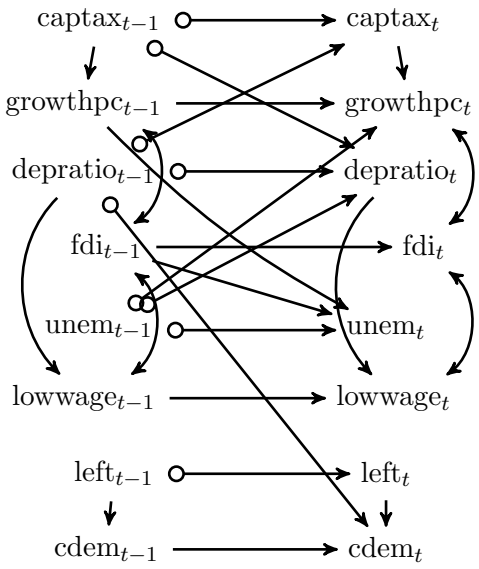


Figure 4: An application of SVAR-FCI ($\alpha \in [0.06, 0.12]$) to the OECD data from Garrett and Mitchell (2001).

studies in international political economy. We follow Beck and Katz (2011) and center the observations at the country level, which is equivalent to forcing country-specific fixed effects. For more details about the data, see the Appendix.

The dependent variable of interest is capital taxation rates (*captax*). The other variables include economic indicators such as GDP growth per capita (*growthpc*) and unemployment (*unem*), variables related to globablization pressures such as total foreign direct investement (*fdi*) and proportion of imports from low-wage countries (*lowwage*), as well as a demographic factor: the ratio of dependents to workers (*depratio*). The data also included two variables to indicate the power of political factions: namely the proportion of cabinet portfolios held by left parties (*left*) and Christian democratic parties (*cdem*).

We used SVAR-FCI with the Fisher Z test of conditional independence to learn a dynamic PAG of order-1, imposing no substantive causal constraints on the specification. As always, in finite samples the resultant model sparsity depends on the tuning parameter α. In Figure 4, we show a search result that was stable over a reasonable range of α values, $\alpha \in [0.06, 0.12]$. At greater than $\alpha = 0.12$, SVAR-FCI recovered successively more edges. We also implemented the data-driven α selection procedure described the Appendix, and present an alternative estimated model therein, where α is selected to be 0.37.

Though we present these results mostly to illustratate the applicability of SVAR-FCI to real data, we note that the results exhibit some interesting features when compared with claims in the literature on this same data set. There is no known "ground truth" for the relationships among these variables, and in fact there is widespread disagreement among empirical studies using the same or similar data, wherein models are specified on the basis of some (controversial) background theory. Garrett and Mitchell (2001), for example, find that some of the strongest determinants of capital taxation rates are unemployment levels,

GDP growth rates, and foreign direct investment. None of these variables, according to our sparse model in Figure 4, are causes of capital taxation rates. In fact, capital taxation rates cause growth. There is also a possible pathway from growth to taxation rates via unemployment and dependency ratio. In the more dense model (Appendix), we see that addition of a directed edge from lagged growth to capital taxation rates, so the taxation and growth processes may exhibit mutual feedback over time. We find some statistical dependencies between taxation rates and variables like FDI and unemployment, but these are best explained by relationships between growth and those factors (some dependencies induced by latent confounders), or possibly explained by the relationship between dependency ratio and unemployment. We find no causal influence from the political faction variables to the rest. Perhaps most interestingly for the background scientific and policy debate, the globablization-related variables (FDI and imports from low-wage countries) show no causal effect on capital taxation rates. Garrett and Mitchell (2001) find at least some weak positive relationship between FDI and taxation rates, whereas in Beck and Katz (2011) they find large regression coefficients for both FDI and low-wage imports in their preferred model specification. In our case, we find that the OECD data supports no causal relationship, even at higher α values: all statistical dependency can be attributed to latent confounding.

6. Discussion

In this paper we have presented two algorithms for learning the structure of dynamic PAGs from multivariate time series, which are asymptotically consistent given appropriate independence tests or model scores. Our simulations suggest that the learning algorithms exhibit high precision but lower recall in finite samples. Of course, decades of algorithm development research in structure learning has produced variations on the basic techniques which improve accuracy in finite samples: potentially, subsampling or stability selection techniques, or adaptive restricted search in score-based case, may increase the recall of SVAR-FCI and SVAR-GFCI by mitigating the propogation of decision errors in finite samples. These improvements are left for future research. Furthermore, in future work we hope to relax the present restriction to acyclic contemporaneous relations and stationary processes.

Acknowledgments

The authors would like to thank David Danks, Clark Glymour, Niels Richard Hansen, Kevin Hoover, Joseph Ramsey, and Kun Zhang. A special thanks is due to Jakob Runge for pointing out errors in an earlier version of this paper. Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative.

Appendix

FCI, GES, and GFCI

In this section we provide psuedocode for FCI, GES, and GFCI to aid in comparison.

Algorithm .1: FCI(Test, α)

Input: Data on variables $\mathbf{X} = \{X_1, ..., X_k\}$
Output: PAG $\mathcal{P}$
1. Form the complete graph $\mathcal{P}$ on vertex set $\mathbf{X}$ with $\circ\!\!-\!\!\circ$ edges.
2. $n \leftarrow 0$
3. **repeat**
4. **for all** pairs of adjacent vertices (X_i, X_j) s.t. $|adj(X_i, \mathcal{P}) \setminus \{X_j\}| \geq n$ and subsets $\mathbf{S} \subset adj(X_i, \mathcal{P}) \setminus \{X_j\}$ s.t. $|\mathbf{S}| = n$
5. **if** $X_i \perp\!\!\!\perp X_j | \mathbf{S}$ according to (Test, α) **then** { Delete edge $X_i \circ\!\!-\!\!\circ X_j$ from $\mathcal{P}$. Let $sepset(X_i, X_j) = sepset(X_j, X_i) = \mathbf{S}$.
6. **end**
7. $n \leftarrow n + 1$
8. **until** for each pair of adjacent vertices (X_i, X_j), $|adj(X_i, \mathcal{P}) \setminus \{X_j\}| < n$.
9. **for all** triples (X_i, X_k, X_j) s.t. $X_i \in adj(X_k, \mathcal{P})$ and $X_j \in adj(X_k, \mathcal{P})$ but $X_i \notin adj(X_j, \mathcal{P})$, orient $X_i \ast\!\!\rightarrow X_k \leftarrow\!\!\ast\, X_j$ iff $X_k \notin sepset(X_i, X_j)$.
10. **for all** pairs (X_i, X_j) adjacent in $\mathcal{P}$ **if** $\exists \mathbf{S}$ s.t. $\mathbf{S} \in pds(X_i, X_j, \mathcal{P})$ or $\mathbf{S} \in pds(X_j, X_i, \mathcal{P})$ and $X_i \perp\!\!\!\perp X_j | \mathbf{S}$ according to (Test, α) **then** { Delete edge $X_i \circ\!\!-\!\!\circ X_j$ from $\mathcal{P}$. Let $sepset(X_i, X_j) = sepset(X_j, X_i) = \mathbf{S}$.
11. Reorient all edges as $\circ\!\!-\!\!\circ$ and **repeat** step 9.
12. Exhaustively apply orientation rules (R1-R10) in Zhang (2008b) to orient remaining endpoints.
13. **return** $\mathcal{P}$.

High-level pseudocode for GES is below. Some of the subroutines are ommitted, see Chickering (2002) or Ramsey et al. (2010) for more details. Specifically, the subroutines ScoreEdgeAddition and ScoreEdgeDeletion determine whether the score increases on adding or deleting the specified edge; ValidInsert and ValidDelete check that some graphical conditions are satisfied; and Rebuild propogates some orientation rules to enforce acylicity. GES can be used in conjunction with any score that is decomposible, score-equivalent, and consistent. The BIC score for multivariate Gaussian data (or multinomial,

in the discrete case) is a popular choice which satisfies these properties. The GES algorithm returns a CPDAG, a representation of an equivalence class of DAGs.

Algorithm .2: GES(Score)

Input: Data on variables $\mathbf{X} = \{X_1, ..., X_k\}$
Output: CPDAG $\mathcal{G}$
1. Form the empty graph $\mathcal{G}$ on vertex set $\mathbf{X}$.
2. Let $S(\mathcal{G}, \mathbf{D})$ be the Score for $\mathcal{G}$ with data $\mathbf{D}$.
3. $\langle \mathcal{G}, S \rangle \leftarrow$ ForwardEquivalenceSearch$(\mathcal{G}, S)$
4. $\mathcal{G} \leftarrow$ BackwardEquivalenceSearch$(\mathcal{G}, S)$
5. **return** $\mathcal{G}$.

Algorithm .3: ForwardEquivalenceSearch$(\mathcal{G}, S)$

Input: Data on variables $\mathbf{X} = \{X_1, ..., X_k\}$
Output: CPDAG $\mathcal{G}$, Score S
1. **while** $\mathbf{E}_0 \neq \emptyset$
2. $\quad \mathbf{E}_0 \leftarrow \mathbf{T}_0 \leftarrow \emptyset$. $S_0 \leftarrow 0$.
3. $\quad$ **for each** edge $E = X_i \rightarrow X_j$ s.t. $X_i \notin adj(X_j, \mathcal{G})$
4. $\quad\quad$ Let $\mathbf{T}' \leftarrow$ vertices X_k s.t. $X_k - X_j$ and $X_k \notin adj(X_i, \mathcal{G})$
5. $\quad\quad$ **for each** subset $\mathbf{T} \in \mathbf{T}'$
6. $\quad\quad\quad$ $\mathcal{G}' \leftarrow$ a DAG in $\mathcal{G}$
7. $\quad\quad\quad$ $S' \leftarrow S +$ ScoreEdgeAddition$(\mathcal{G}, E, \mathbf{T})$
8. $\quad\quad\quad$ **if** $S' > S$ and $S' > S_0$ and ValidInsert$(\mathcal{G}, E, \mathbf{T})$

$$\textbf{then} \begin{cases} \mathbf{E}_0 \leftarrow E \\ \mathbf{T}_0 \leftarrow \mathbf{T} \\ S_0 \leftarrow S' \end{cases}$$

9. $\quad\quad$ **end**
10. $\quad$ **end**
11. $\quad$ **if** $\mathbf{E}_0 \neq \emptyset$

$$\textbf{then} \begin{cases} \text{Add } \mathbf{E}_0 \text{ to } \mathcal{G}. \\ \textbf{for each } T \in \mathbf{T}_0 \textbf{ if } T - X_i \text{ in } \mathcal{G}, \text{ orient } T - X_i \text{ as } T \rightarrow X_i. \\ S \leftarrow S_0 \\ \mathcal{G} \leftarrow \text{Rebuild}(\mathcal{G}) \end{cases}$$

13. **end**
14. **return** $\langle \mathcal{G}, S \rangle$.

Algorithm .4: BackwardEquivalenceSearch($\mathcal{G}, S$)

Input: Data on variables $\mathbf{X} = \{X_1, ..., X_k\}$
Output: CPDAG $\mathcal{G}$
1. **while** $\mathbf{E}_0 \neq \emptyset$
2. $\mathbf{E}_0 \leftarrow \mathbf{H}_0 \leftarrow \emptyset$. $S_0 \leftarrow 0$.
3. **for each** edge E between X_i and X_j in $\mathcal{G}$
4. Let $\mathbf{H}' \leftarrow$ vertices X_k s.t. $X_k - X_j$ and $X_k \in adj(X_i, \mathcal{G})$
5. **for each** subset $\mathbf{H} \in \mathbf{H}'$
6. $\mathcal{G}' \leftarrow$ a DAG in $\mathcal{G}$
7. $S' \leftarrow S +$ ScoreEdgeDeletion$(\mathcal{G}, E, \mathbf{H})$
8. **if** $S' > S$ and $S' > S_0$ and ValidDelete$(\mathcal{G}, E, \mathbf{H})$
then $\begin{cases} \mathbf{E}_0 \leftarrow E \\ \mathbf{H}_0 \leftarrow \mathbf{H} \\ S_0 \leftarrow S' \end{cases}$
9. **end**
10. **end**
11. **if** $\mathbf{E}_0 \neq \emptyset$
then $\begin{cases} \text{Remove } \mathbf{E}_0 \text{ from } \mathcal{G}. \\ \textbf{for each } H \in \mathbf{H}_0 \textbf{ if } X_i - H \text{ in } \mathcal{G}, \text{ orient } X_i - H \text{ as } X_i \rightarrow H. \\ S \leftarrow S_0 \\ \mathcal{G} \leftarrow \text{Rebuild}(\mathcal{G}) \end{cases}$
13. **end**
14. **return** $\mathcal{G}$.

Pseudocode for GFCI, which combines feature of GES with FCI to estimate a PAG, is reproduced below.

Algorithm .5: GFCI(Score, Test, α)

Input: Data on variables $\mathbf{X} = \{X_1, ..., X_k\}$
Output: PAG $\mathcal{P}$
1. $\mathcal{G} \leftarrow$ GES(Score)
2. Form the graph $\mathcal{P}$ on vertex set $\mathbf{X}$ with adjacencies in $\mathcal{G}$ and $\circ\!\!-\!\!\circ$ edges.
3. $n \leftarrow 0$
4. **repeat**
5. **for all** pairs of adjacent vertices (X_i, X_j) s.t. $|adj(X_i, \mathcal{P}) \setminus \{X_j\}| \geq n$ and subsets $\mathbf{S} \subset adj(X_i, \mathcal{P}) \setminus \{X_j\}$ s.t. $|\mathbf{S}| = n$
6. **if** $X_i \perp\!\!\!\perp X_j | \mathbf{S}$ according to (Test, α)
then $\begin{cases} \text{Delete edge } X_i \circ\!\!-\!\!\circ X_j \text{ from } \mathcal{P}. \\ \text{Let } sepset(X_i, X_j) = sepset(X_j, X_i) = \mathbf{S}. \end{cases}$
7. **end**
8. $n \leftarrow n + 1$
9. **until** for each pair of adjacent vertices (X_i, X_j), $|adj(X_i, \mathcal{P}) \setminus \{X_j\}| < n$.
10. **for all** triples (X_i, X_k, X_j) s.t. $X_i \in adj(X_k, \mathcal{P})$ and $X_j \in adj(X_k, \mathcal{P})$ but $X_i \notin adj(X_j, \mathcal{P})$, orient $X_i *\!\!\rightarrow X_k \leftarrow\!\!* X_j$ iff (X_i, X_k, X_j) is a v-structure in $\mathcal{G}$, or it is a triangle in $\mathcal{G}$ and $X_k \notin sepset(X_i, X_j)$.
11. **for all** pairs (X_i, X_j) adjacent in $\mathcal{P}$ **if** $\exists \mathbf{S}$ s.t. $\mathbf{S} \in pds(X_i, X_j, \mathcal{P})$ or $\mathbf{S} \in pds(X_j, X_i, \mathcal{P})$ and $X_i \perp\!\!\!\perp X_j | \mathbf{S}$ according to (Test, α)
then $\begin{cases} \text{Delete edge } X_i \circ\!\!-\!\!\circ X_j \text{ from } \mathcal{P}. \\ \text{Let } sepset(X_i, X_j) = sepset(X_j, X_i) = \mathbf{S}. \end{cases}$
12. Reorient all edges as $\circ\!\!-\!\!\circ$ and **repeat** step 10.
13. Exhaustively apply orientation rules (R1-R10) in Zhang (2008b) to orient remaining endpoints.
14. **return** $\mathcal{P}$.

Data-driven selection of α and p

Maathuis et al. (2009, p. 3144-5) suggest one way of selecting the α tuning parameter required by constraint-based search methods. The idea is to repeat search over a liberal range of α values, and choose the value which leads to a model with the best "fit" as measured by the BIC score. In our case, we estimate a PAG model $\hat{\mathcal{P}}_\alpha$ with SVAR-FCI (or SVAR-GFCI) for some α, and score any MAG in the equivalence class represented by $\hat{\mathcal{P}}_\alpha$ with the BIC score. The likelihood in the BIC score is calculated using ICF (Iterative Conditional Fitting) for MAG models (Drton and Richardson, 2004). Recall that Markov equivalent graphs have the same score, so we can pick any MAG in $\hat{\mathcal{P}}_\alpha$. Then we repeat the procedure for a range of $\alpha \in [a, b]$ with some step size, i.e., we select:

$$\arg\max_{\alpha \in [a,b]} \text{BIC}(\hat{\mathcal{P}}_\alpha)$$

In practice we choose an interval like $[0.01, 0.40]$ with a step size of 0.01. One may follow a similar procedure for selecting the maximum lag length p. Recall that the effective "order" of the marginal process (the maximum lag length p such that $X_{i,t-p}$ is a parent of some $X_{j,t}$ in $\mathcal{P}$) can be different from the true Markov order of the underlying full dynamic DAG, since latent variables induce extra edges in the marginal graph. From the perspective of structure learning and with infinite data, there is no harm in selecting "large" p, since the structure learning method will not add any incorrect edges in the limit. However, with finite data SVAR-FCI or SVAR-GFCI may add extra edges – these would be consistent with the true ancestral relationships, but may reduce the numer of unambiguous orientations in the graph. Furthermore, with a total sample size of T, the effective sample size for all independence tests is $T - p$ so an unnecessarily large p wastes some data. The score-based selection of maximum lag length has a long history in VAR econometrics (see, e.g., Lütkepohl, 2005, p. 148-151). We recommend repeating the above score-maximizing procedure for varying p, i.e., select:

$$\arg\max_{p \in [1,c]} \mathrm{BIC}(\hat{\mathcal{P}}_p)$$

for some maximum considered lag length c and fixed α. One may maximize over both parameters:

$$\arg\max_{\alpha \in [a,b], p \in [1,c]} \mathrm{BIC}(\hat{\mathcal{P}}_{\alpha,p})$$

In each score calculation, the effective sample size in the BIC calculations is set according to the maximum considered lag length, i.e., $T - c$. Of course, maximizing over nontrivial ranges of both α and p can be computationally quite intensive for large models.

Additional details and results on the real data example

As mentioned in the main paper, the data consists of annual observations from 1967 to 1992 of various macroeconomic and political variables on 16 OECD countries. Some countries only report tax rates for a portion of the period under study. Following Beck and Katz (2011), years with missing values in any of the 8 variables were removed from the sample. The missingness occurs mostly at the beginning and end of the data collection period, which is why the data analyzed spans years (at maximum) 1967–1992 though the full data set includes observations from 1961 up to 1994 on some countries. For several countries the first observation is significantly later than 1967, e.g., 1971 (and in one case 1986). Having no missing values "internal" to the study period permits analysis with no imputation or otherwise special handling of missing values, though of course some available data is wasted and the data set is not rectangular. The data can be found online at the following address: `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/27269`

We implemented the α selection procedure described above, considering α values in $[0.01, 0.40]$ by 0.01 increments, and choosing the α which leads to a model with maximimum BIC score in the range. In that case, the procedure selects $\alpha = 0.37$, which we reproduce in Figure 5.

Note that the BIC scores calculated for the resultant models are quite near each other for different values of α, so we doubt there is strong reason to prefer the more dense model. In any case, the different models support roughly the same conclusions with respect to

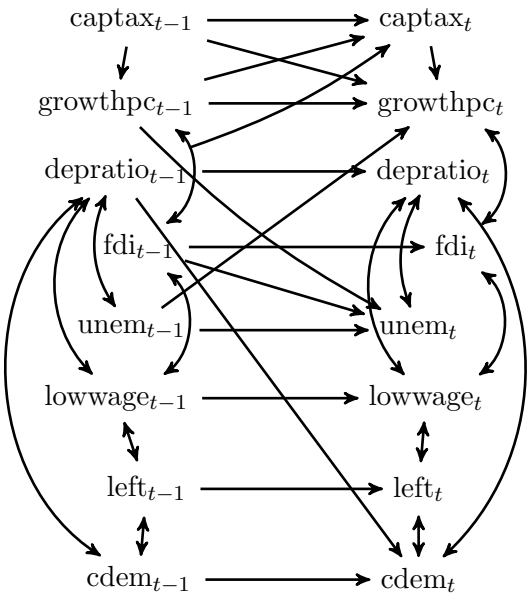


Figure 5: An application of SVAR-FCI ($\alpha = 0.37$, chosen by BIC) to the same OECD data.

capital taxation rates: capital taxation rates are not responding to globalization pressures, at least not very strongly.

References

R Ayesha Ali, Thomas S Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B):2808–2837, 2009.

Nathaniel Beck and Jonathan N Katz. Modeling dynamics in time-series-cross-section political economy data. *Annual Review of Political Science*, 14:331–352, 2011.

David A Bessler and Seongpyo Lee. Money and prices: US data 1869–1914 (a study with directed graphs). *Empirical Economics*, 27(3):427–446, 2002.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

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

Selva Demiralp and Kevin D Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65(s1):745–767, 2003.

Mathias Drton and Thomas Richardson. Iterative conditional fitting for Gaussian ancestral graph models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 130–137. AUAI Press, 2004.

Michael Eichler. Causal inference in time series analysis. In C. Berzuini, A.P. Dawid, and L. Bernardinelli, editors, *Causality: Statistical Perspectives and Applications*. Wiley, 2012.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using FCI. In *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 121–128, 2010.

Franklin M Fisher. A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pages 73–92, 1970.

N Friedman, K Murphy, and S Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 139–147, 1998.

Wei Gao and Zheng Tian. Latent ancestral graph of structure vector autoregressive models. *Journal of Systems Engineering and Electronics*, 21(2):233–238, 2010.

Geoffrey Garrett and Deborah Mitchell. Globalization, government spending and taxation in the OECD. *European Journal of Political Research*, 39(2):145–177, 2001.

Mingming Gong, Kun Zhang, Bernhard Schölkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1898–1906, 2015.

Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

Clive WJ Granger. Economic processes involving feedback. *Information and Control*, 6(1): 28–48, 1963.

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

Clive WJ Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1):199–211, 1988.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, 90:208–225, 2017.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

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

Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A):3133–3164, 2009.

Alessio Moneta, Nadine Chlaß, Doris Entner, and Patrik Hoyer. Causal search in structural vector autoregressive models. In *Proceedings of the NIPS Mini-symposium on Causality in Time Series*, pages 95–114, 2011.

Juan Miguel Ogarrio, Peter Spirtes, and Joseph D Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems*, pages 3303–3311, 2015.

Joseph D Ramsey and Daniel Malinsky. Comparing the performance of graphical structure learning algorithms with TETRAD. *arXiv preprint arXiv:1607.08110*, 2016.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fMRI. *Neuroimage*, 49(2):1545–1558, 2010.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

Pentti Saikkonen. Stability results for nonlinear vector autoregressions with an application to a nonlinear error correction model. Technical report, Discussion Papers of Interdisciplinary Research Project 373 (Humboldt-Universität Berlin), 2001.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 491–498. Morgan Kaufmann Publishers Inc., 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92(437):357–367, 1997.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008a.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008b.