JMLR: Workshop and Conference Proceedings 12 (2011) 95–118 Causality in Time Series

Causal Search in Structural Vector Autoregressive Models

Alessio Moneta MONETA@ECON.MPG.DE
Max Planck Institute of Economics
Jena, Germany

Nadine Chlaß NADINE.CHLASS@UNI-JENA.DE
Friedrich Schiller University of Jena, Germany

Doris Entner DORIS.ENTNER@CS.HELSINKI.FI
Helsinki Institute for Information Technology, Finland

Patrik Hoyer PATRK.HOYER@HELSINKI.FI
Helsinki Institute for Information Technology, Finland

Editor: Florin Popescu and Isabelle Guyon

Abstract

This paper reviews a class of methods to perform causal inference in the framework of a structural vector autoregressive model. We consider three different settings. In the first setting the underlying system is linear with normal disturbances and the structural model is identified by exploiting the information incorporated in the partial correlations of the estimated residuals. Zero partial correlations are used as input of a search algorithm formalized via graphical causal models. In the second, semi-parametric, setting the underlying system is linear with non-Gaussian disturbances. In this case the structural vector autoregressive model is identified through a search procedure based on independent component analysis. Finally, we explore the possibility of causal search in a nonparametric setting by studying the performance of conditional independence tests based on kernel density estimations.

Keywords: Causal inference, econometric time series, SVAR, graphical causal models, independent component analysis, conditional independence tests

1. Introduction

1.1. Causal inference in econometrics

Applied economic research is pervaded by questions about causes and effects. For example, what is the effect of a monetary policy intervention? Is energy consumption causing growth or the other way around? Or does causality run in both directions? Are economic fluctuations mainly caused by monetary, productivity, or demand shocks? Does foreign aid improve living standards in poor countries? Does firms' expenditure in R&D causally influence their profits? Are recent rises in oil prices in part caused by speculation? These are seemingly heterogeneous questions, but they all require some knowledge of the causal process by which variables came to take the values we observe.

A traditional approach to address such questions hinges on the explicit use of *a priori* economic theory. The gist of this approach is to partition a causal process in a deterministic, and a random part and to articulate the deterministic part such as to reflect the causal

dependencies dictated by economic theory. If the formulation of the deterministic part is accurate and reliable enough, the random part is expected to display properties that can easily be analyzed by standard statistical tools. The touchstone of this approach is represented by the work of Haavelmo (1944), which inspired the research program subsequently pursued by the Cowles Commission (Koopmans, 1950; Hood and Koopmans, 1953). Therein, the causal process is formalized by means of a structural equation model, that is, a system of equations with endogenous variables, exogenous variables, and error terms, first developed by Wright (1921). Its coefficients were given a causal interpretation (Pearl, 2000).

This approach has been strongly criticized in the 1970s for being ineffective in both policy evaluation and forecasting. Lucas (1976) pointed out that the economic theory included in the SEM fails to take economic agents' (rational) motivations and expectations into consideration. Agents, according to Lucas, are able to anticipate policy intervention and act contrary to the prediction derived from the structural equation model, since the model usually ignores such anticipations. Sims (1980) puts forth another critique which runs parallel to Lucas' one. It explicitly addresses the status of exogeneity which the Cowles Commission approach attributes (arbitrarily, according to Sims) to some variables such that the structural model can be identified. Sims argues that theory is not a reliable source for deeming a variable as exogenous. More generally, the Cowles Commission approach with its strong *a priori* commitment to theory, risks falling into a vicious circle: if causal information (even if only about direction) can exclusively be derived from background theory, how do we obtain an empirically justified theory? (Cfr. Hoover, 2006, p.75).

An alternative approach has been pursued since Wiener (1956) and Granger's (1969) work. It aims at inferring causal relations directly from the statistical properties of the data relying only to a minimal extent on background knowledge. Granger (1980) proposes a probabilistic concept of causality, similar to Suppes (1970). Granger defines causality in terms of the incremental predictability (at horizon one) of a time series variable $\{Y_t\}$ (given the present and past values of $\{Y_t\}$ and of a set $\{Z_t\}$ of possible relevant variables) when another time series variable $\{X_t\}$ (in its present and past values) is not omitted. More formally:

$$\begin{aligned}&\{X_t\} \text{ Granger-causes } \{Y_t\} \text{ if } P(Y_{t+1}|X_t, X_{t-1}, \ldots, Y_t, Y_{t-1}, \ldots, Z_t, Z_{t-1}, \ldots) \neq \\ &P(Y_{t+1}|Y_t, Y_{t-1}, \ldots, Z_t, Z_{t-1}, \ldots)\end{aligned} \tag{1}$$

As pointed out by Florens and Mouchart (1982), testing the hypothesis of Granger noncausality corresponds to testing conditional independence. Given lags p, $\{X_t\}$ does not Granger cause $\{Y_t\}$, if

$$Y_{t+1} \perp\!\!\!\perp (X_t, X_{t-1}, \ldots, X_{t-p}) \mid (Y_t, Y_{t-1}, \ldots, Y_{t-p}, Z_t, Z_{t-1}, \ldots, Z_{t-p}) \tag{2}$$

To test Granger noncausality, researchers often specify linear vector autoregressive (VAR) models:

$$\mathbf{Y}_t = \mathbf{A}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{3}$$

in which $\mathbf{Y}_t$ is a $k \times 1$ vector of time series variables $(Y_{1,t}, \ldots, Y_{k,t})'$, where $()'$ is the transpose, the $\mathbf{A}_j$ $(j = 1, \ldots, p)$ are $k \times k$ coefficient matrices, and $\mathbf{u}_t$ is the $k \times 1$ vector of random disturbances. In this framework, testing the hypothesis that $\{Y_{i,t}\}$ does not Granger-cause $\{Y_{j,t}\}$, reduces to test whether the (j, i) entries of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$

are vanishing simultaneously. Granger noncausality tests have been extended to nonlinear settings by Baek and Brock (1992), Hiemstra and Jones (1994), and Su and White (2008), using nonparametric tests of conditional independence (more on this topic in section 4).

The concept of Granger causality has been criticized for failing to capture 'structural causality' (Hoover, 2008). Suppose one finds that a variable A Granger-causes another variable B. This does not necessarily imply that an economic mechanism exists by which A can be manipulated to affect B. The existence of such a mechanism in turn does not necessarily imply Granger causality either (for a discussion see Hoover 2001, pp. 150-155). Indeed, the analysis of Granger causality is based on coefficients of reduced-form models, like those incorporated in equation (3), which are unlikely to reliably represent actual economic mechanisms. For instance, in equation (3) the simultaneous causal structure is not modeled in order to facilitate estimation. (However, note that Eichler (2007) and White and Lu (2010) have recently developed and formalized richer structural frameworks in which Granger causality can be fruitfully analyzed.)

1.2. The SVAR framework

Structural vector autoregressive (SVAR) models constitute a middle way between the Cowles Commission approach and the Granger-causality approach. SVAR models aim at recovering the concept of structural causality, but eschew at the same time the strong 'apriorism' of the Cowles Commission approach. The idea is, like in the Cowles Commission approach, to articulate an unobserved structural model, formalized as a dynamic generative model: at each time unit the system is affected by unobserved innovation terms, by which, once filtered by the model, the variables come to take the values we observe. But, differently from the Cowles Commission approach, and similarly to the Granger-VAR model, the data generating process is generally enough articulated so that time series variables are *not* distinguished *a priori* between exogenous and endogenous. A linear SVAR model is in principle a VAR model 'augmented' by the contemporaneous structure:

$$\mathbf{\Gamma}_0 \mathbf{Y}_t = \mathbf{\Gamma}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t. \tag{4}$$

This is easily obtained by pre-multiplying each side of the VAR model

$$\mathbf{Y}_t = \mathbf{A}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{5}$$

by a matrix $\mathbf{\Gamma}_0$ so that $\mathbf{\Gamma}_i = \mathbf{\Gamma}_0 \mathbf{A}_i$, for $i = 1, \ldots, k$ and $\boldsymbol{\varepsilon}_t = \mathbf{\Gamma}_0 \mathbf{u}_t$. Note, however, that not *any* matrix $\mathbf{\Gamma}_0$ will be suitable. The appropriate $\mathbf{\Gamma}_0$ will be that matrix corresponding to the 'right' rotation of the VAR model, that is the rotation compatible both with the contemporaneous causal structure of the variable and the structure of the innovation term. Let us consider a matrix $\mathbf{B}_0 = \mathbf{I} - \mathbf{\Gamma}_0$. If the system is normalized such that the matrix $\mathbf{\Gamma}_0$ has all the elements of the principal diagonal equal to one (which can be done straightforwardly), the diagonal elements of $\mathbf{B}_0$ will be equal to zero. We can write:

$$\mathbf{Y}_t = \mathbf{B}_0 \mathbf{Y}_t + \mathbf{\Gamma}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{\Gamma}_p \mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{6}$$

from which we see that $\mathbf{B}_0$ (and thus $\mathbf{\Gamma}_0$) determines in which form the values of a variable $Y_{i,t}$ will be dependent on the contemporaneous value of another variable $Y_{j,t}$. The 'right'

rotation will also be the one which makes $\boldsymbol{\varepsilon}_t$ a vector of authentic innovation terms, which are expected to be independent (not only over time, but also contemporaneously) sources or shocks.

In the literature, different methods have been proposed to identify the SVAR model (4) on the basis of the estimation of the VAR model (5). Notice that there are more unobserved parameters in (4), whose number amounts to $k^2(p+1)$, than parameters that can be estimated from (5), which are $k^2p + k(k+1)/2$, so one has to impose at least $k(k-1)/2$ restrictions on the system. One solution to this problem is to get a rotation of (5) such that the covariance matrix of the SVAR residuals $\boldsymbol{\Sigma}_\varepsilon$ is diagonal, using the Cholesky factorization of the estimated residuals $\boldsymbol{\Sigma}_{\mathbf{u}}$. That is, let $\mathbf{P}$ be the lower-triangular Cholesky factorization of $\boldsymbol{\Sigma}_{\mathbf{u}}$ (i.e. $\boldsymbol{\Sigma}_{\mathbf{u}} = \mathbf{P}\mathbf{P}'$), let $\mathbf{D}$ be a $k \times k$ diagonal matrix with the same diagonal as $\mathbf{P}$, and let $\boldsymbol{\Gamma}_0 = \mathbf{D}\mathbf{P}^{-1}$. By pre-multiplying (5) by $\boldsymbol{\Gamma}_0$, it turns out that $\boldsymbol{\Sigma}_\varepsilon = E[\boldsymbol{\Gamma}_0\mathbf{u}_t\mathbf{u}_t'\boldsymbol{\Gamma}_0'] = \mathbf{D}\mathbf{D}'$, which is diagonal. A problem with this method is that $\mathbf{P}$ changes if the ordering of the variables $(Y_{1t}, \ldots, Y_{kt})'$ in $\mathbf{Y}_t$ and, consequently, the order of residuals in $\boldsymbol{\Sigma}_{\mathbf{u}}$, changes. Since researchers who estimate a SVAR are often exclusively interested on tracking down the effect of a structural shock ε_{it} on the variables $Y_{1,t}, \ldots, Y_{k,t}$ over time (*impulse response functions*), Sims (1981) suggested investigating to what extent the impulse response functions remain robust under changes of the order of variables.

Popular alternatives to the Cholesky identification scheme are based either on the use of *a priori*, theory-based, restrictions or on the use of long-run restrictions. The former solution consists in imposing economically plausible constraints on the contemporaneous interactions among variables (Blanchard and Watson, 1986; Bernanke, 1986) and has the drawback of ultimately depending on the *a priori* reliability of economic theory, similarly to the Cowles Commission approach. The second solution is based on the assumptions that certain economic shocks have long-run effect to other variables, but do not influence in the long-run the level of other variables (see Shapiro and Watson, 1988; Blanchard and Quah, 1989; King et al., 1991). This approach has been criticized as not being very reliable unless strong *a priori* restrictions are imposed (see Faust and Leeper, 1997).

In the rest of the paper, we first present a method, based on the graphical causal model framework, to identify the SVAR (section 2). This method is based on conditional independence tests among the estimated residuals of the VAR estimated model. Such tests rely on the assumption that the shocks affecting the model are Gaussian. We then relax the Gaussianity assumption and present a method to identify the SVAR model based on independent component analysis (section 3). Here the main assumption is that shocks are non-Gaussian and independent. Finally (section 4), we explore the possibility of extending the framework for causal inference to a nonparametric setting. In section 5 we wrap up the discussion and conclude by formulating some open questions.

2. SVAR identification via graphical causal models

2.1. Background

A data-driven approach to identify the structural VAR is based on the analysis of the estimated residuals $\hat{\mathbf{u}}_t$. Notice that when a basic VAR model is estimated (equation 3), the information about contemporaneous causal dependence is incorporated exclusively in the

residuals (being not modeled among the variables). Graphical causal models, as originally developed by Pearl (2000) and Spirtes et al. (2000), represent an efficient method to recover, at least in part, the contemporaneous causal structure moving from the analysis of the conditional independencies among the estimated residuals. Once the contemporaneous causal structure is recovered, the estimation of the lagged autoregressive coefficients permits us to identify the complete SVAR model.

This approach was initiated by Swanson and Granger (1997), who proposed to test whether a particular causal order of the VAR is in accord with the data by testing all the partial correlations of order one among error terms and checking whether some partial correlations are vanishing. Reale and Wilson (2001), Bessler and Lee (2002), Demiralp and Hoover (2003), and Moneta (2008) extended the approach by using the partial correlations of the VAR residuals as input to graphical causal model search algorithms.

In graphical causal models, the structural model is represented as a causal graph (a *Directed Acyclic Graph* if the presence of causal loops is excluded), in which each node represents a random variable and each edge a causal dependence. Furthermore, a set of assumptions or 'rules of inference' are formulated, which regulate the relationship between causal and probabilistic dependencies: the *causal Markov* and the *faithfulness* conditions (Spirtes et al., 2000). The former restricts the joint probability distribution of modeled variables: each variable is independent of its graphical non-descendants conditional on its graphical parents. The latter makes causal discovery possible: all of the conditional independence relations among the modeled variables follow from the causal Markov condition. Thus, for example, if the causal structure is represented as $Y_{1t} \rightarrow Y_{2t} \rightarrow Y_{t,3}$, it follows from the Markov condition that $Y_{1,t} \perp\!\!\!\perp Y_{3,t}|Y_{2,t}$. If, on the other hand, the only (conditional) independence relation among $Y_{1,t}, Y_{2,t}, Y_{3,t}$ is $Y_{1,t} \perp\!\!\!\perp Y_{3,t}$, it follows from the faithfulness condition that $Y_{1,t} \rightarrow Y_{3,t} <\!\!- Y_{2,t}$.

Constraint-based algorithms for causal discovery, like for instance, PC, SGS, FCI (Spirtes et al. 2000), or CCD (Richardson and Spirtes 1999), use tests of conditional independence to constrain the possible causal relationships among the model variables. The first step of the algorithm typically involves the formation of a complete undirected graph among the variables so that they are all connected by an undirected edge. In a second step, conditional independence relations (or d-separations, which are the graphical characterization of conditional independence) are merely used to erase edges and, in further steps, to direct edges. The output of such algorithms are not necessarily one single graph, but a class of *Markov equivalent* graphs.

There is nothing neither in the Markov or faithfulness condition, nor in the constraint-based algorithms that limits them to linear and Gaussian settings. Graphical causal models do not require *per se* any a priori specification of the functional dependence between variables. However, in applications of graphical models to SVAR, conditional independence is ascertained by testing vanishing partial correlations (Swanson and Granger, 1997; Bessler and Lee, 2002; Demiralp and Hoover, 2003; Moneta, 2008). Since normal distribution guarantees the equivalence between zero partial correlation and conditional independence, these applications deal *de facto* with linear and Gaussian processes.

2.2. Testing residuals zero partial correlations

There are alternative methods to test zero partial correlations among the error terms $\hat{\mathbf{u}}_t = (u_{1t}, \ldots, u_{kt})'$. Swanson and Granger (1997) use the partial correlation coefficient. That is, in order to test, for instance, $\rho(u_{it}, u_{kt}|u_{jt}) = 0$, they use the standard t statistics from a least square regression of the model:

$$u_{it} = \alpha_j u_{jt} + \alpha_k u_{kt} + \varepsilon_{it}, \tag{7}$$

on the basis that $\alpha_k = 0 \Leftrightarrow \rho(u_{it}, u_{kt}|u_{jt}) = 0$. Since Swanson and Granger (1997) impose the partial correlation constraints looking only at the set of partial correlations of order one (that is conditioned on only one variable), in order to run their tests they consider regression equations with only two regressors, as in equation (7).

Bessler and Lee (2002) and Demiralp and Hoover (2003) use Fisher's z that is incorporated in the software TETRAD (Scheines et al., 1998):

$$z(\rho_{XY.\mathbf{K}}, T) = \frac{1}{2}\sqrt{T - |\mathbf{K}| - 3}\ \log\left(\frac{|1 + \rho_{XY.\mathbf{K}}|}{|1 - \rho_{XY.\mathbf{K}}|},\right) \tag{8}$$

where $|\mathbf{K}|$ equals the number of variables in $\mathbf{K}$ and T the sample size. If the variables (for instance $X = u_{it}$, $Y = u_{kt}$, $\mathbf{K} = (u_{jt}, u_{ht})$) are normally distributed, we have that

$$z(\rho_{XY.\mathbf{K}}, T) - z(\hat{\rho}_{XY.\mathbf{K}}, T) \sim N(0, 1) \tag{9}$$

(see Spirtes et al., 2000, p.94).

A different approach, which takes into account the fact that correlations are obtained from residuals of a regression, is proposed by Moneta (2008). In this case it is useful to write the VAR model of equation (3) in a more compact form:

$$\mathbf{Y}_t = \mathbf{\Pi}'\mathbf{X}_t + \mathbf{u}_t, \tag{10}$$

where $\mathbf{X}'_t = [\mathbf{Y}'_{t-1}, ..., \mathbf{Y}'_{t-p}]$, which has dimension $(1 \times kp)$ and $\mathbf{\Pi}' = [\mathbf{A}_1, \ldots, \mathbf{A}_p]$, which has dimension $(k \times kp)$. In case of stable VAR process (see next subsection), the conditional maximum likelihood estimate of Π for a sample of size T is given by

$$\hat{\mathbf{\Pi}}' = \left[\sum_{t=1}^{T} \mathbf{Y}_t \mathbf{X}'_t\right] \left[\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}'_t\right]^{-1}.$$

Moreover, the ith row of $\hat{\Pi}'$ is

$$\hat{\pi}'_i = \left[\sum_{t=1}^{T} Y_{it} \mathbf{X}'_t\right] \left[\sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}'_t\right]^{-1},$$

which coincides with the estimated coefficient vector from an OLS regression of Y_{it} on $\mathbf{X}_t$ (Hamilton 1994: 293). The maximum likelihood estimate of the matrix of variance and covariance among the error terms $\mathbf{\Sigma}_u$ turns out to be $\hat{\mathbf{\Sigma}}_u = (1/T)\sum_{t=1}^{T} \hat{\mathbf{u}}_t \hat{\mathbf{u}}'_t$, where $\hat{\mathbf{u}}_t = \mathbf{Y}_t - \hat{\mathbf{\Pi}}'\mathbf{X}_t$. Therefore, the maximum likelihood estimate of the covariance between u_{it}

and u_{jt} is given by the (i,j) element of $\hat{\mathbf{\Sigma}}_u$: $\hat{\sigma}_{ij} = (1/T)\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt}$. Denoting by σ_{ij} the (i,j) element of $\mathbf{\Sigma}_u$, let us first define the following matrix transform operators: *vec*, which stacks the columns of a $k \times k$ matrix into a vector of length k^2 and *vech*, which vertically stacks the elements of a $k \times k$ matrix on or below the principal diagonal into a vector of length $k(k+1)/2$. For example:

$$\text{vec}\begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix}, \quad \text{vech}\begin{bmatrix} \sigma_{11}\sigma_{12} \\ \sigma_{21}\sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \end{bmatrix}.$$

The process being stationary and the error terms Gaussian, it turns out that:

$$\sqrt{T}\,[\text{vech}(\hat{\mathbf{\Sigma}}_u) - \text{vech}(\mathbf{\Sigma}_u)] \xrightarrow{d} N(\mathbf{0},\ \mathbf{\Omega}), \tag{11}$$

where $\mathbf{\Omega} = 2\mathbf{D}_k^{+}(\mathbf{\Sigma}_u \otimes \mathbf{\Sigma}_u)(\mathbf{D}_k^{+})'$, $\mathbf{D}_k^{+} \equiv (\mathbf{D}_k'\mathbf{D}_k)^{-1}\mathbf{D}_k'$, $\mathbf{D}_k$ is the unique $(k^2 \times k(k+1)/2)$ matrix satisfying $\mathbf{D}_k\text{vech}(\mathbf{\Omega}) = \text{vec}(\mathbf{\Omega})$, and $\otimes$ denotes the Kronecker product (see Hamilton 1994: 301). For example, for $k = 2$, we have,

$$\sqrt{T}\begin{bmatrix} \hat{\sigma}_{11} - \sigma_{11} \\ \hat{\sigma}_{12} - \sigma_{12} \\ \hat{\sigma}_{22} - \sigma_{22} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\sigma_{11}^2 & 2\sigma_{11}\sigma_{12} & 2\sigma_{12}^2 \\ 2\sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 & 2\sigma_{12}\sigma_{22} \\ 2\sigma_{12}^2 & 2\sigma_{12}\sigma_{22} & 2\sigma_{22}^2 \end{bmatrix}\right)$$

Therefore, to test the null hypothesis that $\rho(u_{it}, u_{jt}) = 0$ from the VAR estimated residuals, it is possible to use the Wald statistic:

$$\frac{T\,(\hat{\sigma}_{ij})^2}{\hat{\sigma}_{ii}\hat{\sigma}_{jj} + \hat{\sigma}_{ij}^2} \approx \chi^2(1).$$

The Wald statistic for testing vanishing partial correlations of any order is obtained by applying the delta method, which suggests that if X_T is a $(r \times 1)$ sequence of vector-valued random-variables and if $[\sqrt{T}(X_{1T} - \theta_1), \ldots, \sqrt{T}(X_{rT} - \theta_r)] \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and $h_1, \ldots, h_r$ are r real-valued functions of $\theta = (\theta_1, \ldots, \theta_r)$, $h_i : \mathbf{R}^r \to \mathbf{R}$, defined and continuously differentiable in a neighborhood ω of the parameter point θ and such that the matrix $B = ||\partial h_i/\partial\theta_j||$ of partial derivatives is nonsingular in ω, then:

$$[\sqrt{T}[h_1(X_T) - h_1(\theta)], \ldots, \sqrt{T}[h_r(X_T) - h_r(\theta)]] \xrightarrow{d} N(\mathbf{0}, B\Sigma B')$$

(see Lehmann and Casella 1998: 61).

Thus, for $k = 4$, suppose one wants to test $corr(u_{1t}, u_{3t}|u_{2t}) = 0$. First, notice that $\rho(u_1, u_3|u_2) = 0$ if and only if $\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23} = 0$ (by definition of partial correlation). One can define a function $g : \mathbf{R}^{k(k+1)/2} \to \mathbf{R}$, such that $g(vech(\Sigma_u)) = \sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}$. Thus,

$$\nabla g' = (0,\ -\sigma_{23},\ \sigma_{22},\ 0,\ \sigma_{13},\ -\sigma_{12},\ 0,\ 0,\ 0,\ 0).$$

Applying the delta method:

$$\sqrt{T}[(\hat{\sigma}_{22}\hat{\sigma}_{13} - \hat{\sigma}_{12}\hat{\sigma}_{23}) - (\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23})] \xrightarrow{d} N(0, \nabla g'\mathbf{\Omega}\nabla g).$$

The Wald test of the null hypothesis $corr(u_{1t}, u_{3t}|u_{2t}) = 0$ is given by:

$$\frac{T(\hat{\sigma}_{22}\hat{\sigma}_{13} - \hat{\sigma}_{12}\hat{\sigma}_{23})^2}{\nabla g' \Omega \nabla g} \approx \chi^2(1).$$

Tests for higher order correlations and for $k > 4$ follow analogously (see also Moneta, 2003). This testing procedure has the advantage, with respect to the alternative methods, to be straightforwardly applied to the case of cointegrated data, as will be explained in the next subsection.

2.3. Cointegration case

A typical feature of economic time series data in which there is some form of causal dependence is cointegration. This term denotes the phenomenon that nonstationary processes can have linear combinations that are stationary. That is, suppose that each component Y_{it} of $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{kt})'$, which follows the VAR process

$$\mathbf{Y}_t = \mathbf{A}_1 \mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{u}_t,$$

is nonstationary and integrated of order one ($\sim I(1)$). This means that the VAR process $\mathbf{Y}_t$ is not *stable*, i.e. $\det(I_k - A_1 z - A_p z^p)$ is equal to zero for some $|z| \leq 1$ (Lütkepohl, 2006), and that each component ΔY_{it} of $\Delta \mathbf{Y}_t = (\mathbf{Y}_t - \mathbf{Y}_{t-1})$ is stationary ($I(0)$), that is it has time-invariant means, variances and covariance structure. A linear combination between between the elements of $\mathbf{Y}_t$ is called a *cointegrating relationship* if there is a linear combination $c_1 Y_{1t} + \ldots + c_k Y_{kt}$ which is stationary ($I(0)$).

If it is the case that the VAR process is unstable with the presence of cointegrating relationships, it is more appropriate (Lütkepohl, 2006; Johansen, 2006) to estimate the following re-parametrization of the VAR model, called Vector Error Correction Model (VECM):

$$\Delta \mathbf{Y}_t = \mathbf{F}_1 \Delta \mathbf{Y}_{t-1} + \ldots + \mathbf{F}_{p-1} \Delta \mathbf{Y}_{t-p+1} - \mathbf{G}\mathbf{Y}_{t-p} + \mathbf{u}_t, \tag{12}$$

where $\mathbf{F}_i = -(\mathbf{I}_k - \mathbf{A}_1 - \ldots - \mathbf{A}_i)$, for $i = 1, \ldots, p-1$ and $\mathbf{G} = \mathbf{I}_k - \mathbf{A}_1 - \ldots - \mathbf{A}_p$. The $(k \times k)$ matrix $\mathbf{G}$ has rank r and thus $\mathbf{G}$ can be written as $\mathbf{HC}$ with $\mathbf{H}$ and $\mathbf{C}'$ of dimension $(k \times r)$ and of rank r. $\mathbf{C} \equiv [c_1, \ldots, c_r]'$ is called the *cointegrating matrix*.

Let $\tilde{\mathbf{C}}, \tilde{\mathbf{H}}$, and $\tilde{\mathbf{F}}_i$ be the maximum likelihood estimator of $\mathbf{C}$, $\mathbf{H}$, $\mathbf{F}$ according to Johansen's (1988, 1991) approach. Then the asymptotic distribution of $\tilde{\mathbf{\Sigma}}_u$, that is the maximum likelihood estimator of the covariance matrix of u_t, is:

$$\sqrt{T}\,[\mathrm{vech}(\tilde{\mathbf{\Sigma}}_u) - \mathrm{vech}(\mathbf{\Sigma}_u)] \xrightarrow{d} N(\mathbf{0},\ 2\mathbf{D}_k^+(\mathbf{\Sigma}_u \otimes \mathbf{\Sigma}_u)\mathbf{D}_k^{+\prime}), \tag{13}$$

which is equivalent to equation (11) (see it again for the definition of the various operators). Thus, it turns out that the asymptotic distribution of the maximum likelihood estimator $\tilde{\mathbf{\Sigma}}_u$ is the same as the OLS estimation $\hat{\mathbf{\Sigma}}_u$ for the case of stable VAR.

Thus, the application of the method described for testing residuals zero partial correlations can be applied straightforwardly to cointegrated data. The model is estimated as a VECM error correction model using Johansen's (1988, 1991) approach, correlations are tested exploiting the asymptotic distribution of $\tilde{\Sigma}_u$ and finally can be parameterized back in its VAR form of equation (3).

2.4. Summary of the search procedure

The graphical causal models approach to SVAR identification, which we suggest in case of Gaussian and linear processes, can be summarized as follows.

Step 1 Estimate the VAR model $\mathbf{Y}_t = \mathbf{A}_1\mathbf{Y}_{t-1} + \ldots + \mathbf{A}_p\mathbf{Y}_{t-p} + \mathbf{u}_t$ with the usual specification tests about normality, zero autocorrelation of residuals, lags, and unit roots (see Lütkepohl, 2006). If the hypothesis of nonstationarity is rejected, estimate the VAR model via OLS (equivalent to MLE under the assumption of normality of the errors). If unit root tests do not reject $I(1)$ nonstationarity in the data, specify the model as VECM testing the presence of cointegrating relationships. If tests suggest the presence of cointegrating relationships, estimate the model as VECM. If cointegration is rejected estimate the VAR models taking first difference.

Step 2 Run tests for zero partial correlations between the elements $u_{1t}, \ldots, u_{kt}$ of $\mathbf{u}_t$ using the Wald statistics on the basis of the asymptotic distribution of the covariance matrix of $\mathbf{u}_t$. Note that not all possible partial correlations $\rho(u_{it}, u_{jt}|u_{ht}, \ldots)$ need to be tested, but only those necessary for step 3.

Step 3 Apply a causal search algorithm to recover the causal structure among $u_{1t}, \ldots, u_{kt}$, which is equivalent to the causal structure among $Y_{1t}, \ldots, Y_{kt}$ (cfr. section 1.2 and see Moneta 2003). In case of acyclic (no feedback loops) and causally sufficient (no latent variables) structure, the suggested algorithm is the PC algorithm of Spirtes et al. (2000, pp. 84-85). Moneta (2008) suggested few modifications to the PC algorithm in order to make the orientation of edges compatible with as many conditional independence tests as possible. This increases the computational time of the search algorithm, but considering the fact that VAR models deal with a few number of time series variables (rarely more than six to eight; see Bernanke et al. 2005), this slowing down does not create a serious concern in this context. Table 1 reports the modified PC algorithm. In case of acyclic structure without causal sufficiency (i.e. possibly including latent variables), the suggested algorithm is FCI (Spirtes et al. 2000, pp. 144-145). In the case of no latent variables and in the presence of feedback loops, the suggested algorithm is CCD (Richardson and Spirtes, 1999). There is no algorithm in the literature which is consistent for search when both latent variables and feedback loops may be present. If the goal of the study is only impulse response analysis (i.e. tracing out the effects of structural shocks $\varepsilon_{1t}, \ldots, \varepsilon_{kt}$ on $\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots$) and neither contemporaneous feedbacks nor latent variables can be excluded *a priori*, a possible solution is to apply only steps (A) and (B) of the PC algorithm. If the resulting set of possible causal structures (represented by an undirected graph) contains a manageable number of elements, one can study the characteristics of the impulse response functions which are robust across all the possible causal structures, where the presence of both feedbacks and latent variables is allowed (Moneta, 2004).

Step 4 Calculate structural coefficients and impulse response functions. If the output of Step 3 is a set of causal structures, run sensitivity analysis to investigate the robustness of the conclusions under the different possible causal structures. Bootstrap procedures may

also be applied to determine which is the most reliable causal order (see simulations and applications in Demiralp et al., 2008).

3. Identification via independent component analysis

The methods considered in the previous section use tests for zero partial correlation on the VAR-residuals to obtain (partial) information about the contemporaneous structure in an SVAR model with Gaussian shocks. In this section we show how non-Gaussian and independent shocks can be exploited for model identification by using the statistical method of 'Independent Component Analysis' (ICA, see Comon (1994); Hyvärinen et al. (2001)). The method is again based on the VAR-residuals $\mathbf{u}_t$ which can be obtained as in the Gaussian case by estimating the VAR model using for example ordinary least squares or least absolute deviations, and can be tested for non-Gaussianity using any normality test (such as the Shapiro-Wilk or Jarque-Bera test).

To motivate, we note that, from equations (3) and (4) (with matrix $\mathbf{\Gamma}_0$) or the Cholesky factorization in section 1.2 (with matrix $\mathbf{PD}^{-1}$), the VAR-disturbances $\mathbf{u}_t$ and the structural shocks $\boldsymbol{\varepsilon}_t$ are connected by

$$\mathbf{u}_t = \mathbf{\Gamma}_0^{-1}\boldsymbol{\varepsilon}_t = \mathbf{PD}^{-1}\boldsymbol{\varepsilon}_t \tag{14}$$

with square matrices $\mathbf{\Gamma}_0$ and $\mathbf{PD}^{-1}$, respectively. Equation (14) has two important properties: First, the vectors $\mathbf{u}_t$ and $\boldsymbol{\varepsilon}_t$ are of the same length, meaning that there are as many residuals as structural shocks. Second, the residuals $\mathbf{u}_t$ are linear mixtures of the shocks $\boldsymbol{\varepsilon}_t$, connected by the 'mixing matrix' $\mathbf{\Gamma}_0^{-1}$. This resembles the ICA model, when placing certain assumptions on the shocks $\boldsymbol{\varepsilon}_t$.

In short, the ICA model is given by $\mathbf{x} = \mathbf{As}$, where $\mathbf{x}$ are the mixed components, $\mathbf{s}$ the independent, non-Gaussian sources, and $\mathbf{A}$ a square invertible mixing matrix (meaning that there are as many mixtures as independent components). Given samples from the mixtures $\mathbf{x}$, ICA estimates the mixing matrix $\mathbf{A}$ and the independent components $\mathbf{s}$, by linearly transforming $\mathbf{x}$ in such a way that the dependencies among the independent components $\mathbf{s}$ are minimized. The solution is unique up to ordering, sign and scaling (Comon, 1994; Hyvärinen et al., 2001).

By comparing the ICA model $\mathbf{x} = \mathbf{As}$ and equation (14), we see a one-to-one correspondence of the mixtures $\mathbf{x}$ to the residuals $\mathbf{u}_t$ and the independent components $\mathbf{s}$ to the shocks $\boldsymbol{\varepsilon}_t$. Thus, to be able to apply ICA, we need to assume that the shocks are non-Gaussian and mutually independent. We want to emphasize that no specific non-Gaussian distribution is assumed for the shocks, but only that they cannot be Gaussian.[1] For the shocks to be mutually independent their joint distribution has to factorize into the product of the marginal distributions. In the non-Gaussian setting, this implies zero partial correlation, but the converse is not true (as opposed to the Gaussian case where the two statements are equivalent). Thus, for non-Gaussian distributions conditional independence is a much stronger requirement than uncorrelatedness.

Under the assumption that the shocks $\boldsymbol{\varepsilon}_t$ are non-Gaussian and independent, equation (14) follows exactly the ICA-model and applying ICA to the VAR residuals $\mathbf{u}_t$ yields a unique solution (up to ordering, sign, and scaling) for the mixing matrix $\mathbf{\Gamma}_0^{-1}$ and the

1. Actually, the requirement is that *at most one* of the residuals can be Gaussian.

Table 1: Search algorithm (adapted from the PC Algorithm of Spirtes et al. (2000: 84-85); in bold character the modifications).

Under the assumption of Gaussianity conditional independence is tested by zero partial correlation tests.

(A): (*connect everything*):
Form the complete undirected graph $\mathcal{G}$ on the vertex set $u_{1t}, \ldots, u_{kt}$ so that each vertex is connected to any other vertex by an undirected edge.
(B)(*cut some edges*):
$n = 0$
repeat :
    repeat :
        select an ordered pair of variables u_{ht} and u_{it} that are adjacent in $\mathcal{G}$ such that the number of variables adjacent to u_{ht} is equal or greater than $n + 1$. Select a set S of n variables adjacent to u_{ht} such that $u_{ti} \notin S$. If $u_{ht} \perp\!\!\!\perp u_{it}|S$ delete edge u_{ht} — u_{it} from $\mathcal{G}$;
    until all ordered pairs of adjacent variables u_{ht} and u_{it} such that the number of variables adjacent to u_{ht} is equal or greater than $n + 1$ and all sets S of n variables adjacent to u_{ht} such that $u_{it} \notin S$ have been checked to see if $u_{ht} \perp\!\!\!\perp u_{it}|S$;
    $n = n + 1$;
until for each ordered pair of adjacent variables u_{ht}, u_{it}, the number of adjacent variables to u_{ht} is less than $n + 1$;
(C)(*build colliders*):
for each triple of vertices u_{ht}, u_{it}, u_{jt} such that the pair u_{ht}, u_{it} and the pair u_{it}, u_{jt} are each adjacent in $\mathcal{G}$ but the pair u_{ht}, u_{jt} is not adjacent in $\mathcal{G}$, orient u_{ht} — u_{it} — u_{jt} as $u_{ht} \longrightarrow u_{it} <\!\!-\!\!- u_{jt}$ if and only if u_{it} does not belong to **any set** of variables S such that $u_{ht} \perp\!\!\!\perp u_{jt}|S$;
(D)(*direct some other edges*):
repeat :
    if $u_{at} \longrightarrow u_{bt}$, u_{bt} and u_{ct} are adjacent, u_{at} and u_{ct} are not adjacent and u_{bt} belongs to **every set** S such that $u_{at} \perp\!\!\!\perp u_{ct}|S$, then orient u_{bt} — u_{ct} as $u_{bt} \longrightarrow u_{ct}$;
    if there is a directed path from u_{at} to u_{bt}, and an edge between u_{at} and u_{bt}, then orient u_{at} — u_{bt} as $u_{at} \longrightarrow u_{bt}$;
until no more edges can be oriented.

independent components $\boldsymbol{\varepsilon}_t$ (i.e. the structural shocks in our case). However, the ambiguities of ICA make it hard to directly interpret the shocks found by ICA since without further analysis we cannot relate the shocks directly to the measured variables.

Hence, we assume that the residuals $\mathbf{u}_t$ follow a linear non-Gaussian acyclic model (Shimizu et al., 2006), which means that the contemporaneous structure is represented by a DAG (directed acyclic graph). In particular, the model is given by

$$\mathbf{u}_t = \mathbf{B}_0 \mathbf{u}_t + \boldsymbol{\varepsilon}_t \tag{15}$$

with a matrix $\mathbf{B}_0$, whose diagonal elements are all zero and, if permuted according to the causal order, is strictly lower triangular. By rewriting equation (15) we see that

$$\boldsymbol{\Gamma}_0 = \mathbf{I} - \mathbf{B}_0. \tag{16}$$

From this equation it follows that the matrix $\mathbf{B}_0$ describes the contemporaneous structure of the variables $\mathbf{Y}_t$ in the SVAR model as shown in equation (6). Thus, if we can identify the matrix $\boldsymbol{\Gamma}_0$, we also obtain the matrix $\mathbf{B}_0$ for the contemporaneous effects. As pointed out above, the matrix $\boldsymbol{\Gamma}_0^{-1}$ (and hence $\boldsymbol{\Gamma}_0$) can be estimated using ICA up to ordering, scaling, and sign. With the restriction of $\mathbf{B}_0$ representing an acyclic system, we can resolve these ambiguities and are able to fully identify the model. For simplicity, let us assume that the variables are arranged according to a causal ordering, so that the matrix $\mathbf{B}_0$ is strictly lower triangular. From equation (16) then follows that the matrix $\boldsymbol{\Gamma}_0$ is lower triangular with all ones on the diagonal. Using this information, the ambiguities of ICA can be resolved in the following way.

The lower triangularity of $\mathbf{B}_0$ allows us to find the unique permutation of the rows of $\boldsymbol{\Gamma}_0$, which yields all non-zero elements on the diagonal of $\boldsymbol{\Gamma}_0$, meaning that we replace the matrix $\boldsymbol{\Gamma}_0$ with $\mathbf{Q}_1\boldsymbol{\Gamma}_0$ where $\mathbf{Q}_1$ is the uniquely determined permutation matrix. Finding this permutation resolves the ordering-ambiguity of ICA and links the shocks $\boldsymbol{\varepsilon}_t$ to the components of the residuals $\mathbf{u}_t$ in a one-to-one manner. The sign- and scaling-ambiguity is now easy to fix by simply dividing each row of $\boldsymbol{\Gamma}_0$ (the row-permuted version from above) by the corresponding diagonal element yielding all ones on the diagonal, as implied by Equation (16). This ensures that the connection strength of the shock $\boldsymbol{\varepsilon}_t$ on the residual $\mathbf{u}_t$ is fixed to one in our model (Equation (15)).

For the general case where $\mathbf{B}_0$ is not arranged in the causal order, the above arguments for solving the ambiguities still apply. Furthermore, we can find the causal order of the contemporaneous variables by performing simultaneous row- and column-permutations on $\boldsymbol{\Gamma}_0$ yielding the matrix closest to lower triangular, in particular $\tilde{\boldsymbol{\Gamma}}_0 = \mathbf{Q}_2\boldsymbol{\Gamma}_0\mathbf{Q}_2'$ with an appropriate permutation matrix $\mathbf{Q}_2$. In case non of these permutations leads to a close to lower triangular matrix a warning is issued.

Essentially, the assumption of acyclicity allows us to uniquely connect the structural shocks $\boldsymbol{\varepsilon}_t$ to the components of $\mathbf{u}_t$ and fully identify the contemporaneous structure. Details of the procedure can be found in (Shimizu et al., 2006; Hyvärinen et al., 2010). In the sense of the Cholesky factorization of the covariance matrix explained in Section 1 (with $\mathbf{P}\mathbf{D}^{-1} = \boldsymbol{\Gamma}_0^{-1}$), full identifiability means that a causal order among the contemporaneous variables can be determined.

In addition to yielding full identification, an additional benefit of using the ICA-based procedure when shocks are non-Gaussian is that it does not rely on the faithfulness assumption, which was necessary in the Gaussian case.

We note that there are many ways of exploiting non-Gaussian shocks for model identification as alternatives to directly using ICA. One such approach was introduced by Shimizu et al. (2009). Their method relies on iteratively finding an exogenous variable and regressing out their influence on the remaining variables. An exogenous variable is characterized by being independent of the residuals when regressing any other variable in the model on it. Starting from the model in equation (15), this procedure returns a causal ordering of the variables $\mathbf{u}_t$ and then the matrix $\mathbf{B}_0$ can be estimated using the Cholesky approach.

One relatively strong assumption of the above methods is the *acyclicity* of the contemporaneous structure. In (Lacerda et al., 2008) an extension was proposed where feedback loops were allowed. In terms of the matrix $\mathbf{B}_0$ this means that it is not restricted to being lower triangular (in an appropriate ordering of the variables). While in general this model is not identifiable because we cannot uniquely match the shocks to the residuals, Lacerda et al. (2008) showed that the model is identifiable when assuming stability of the generating model in (15) (the absolute value of the biggest eigenvalue in $\mathbf{B}_0$ is smaller than one) and disjoint cycles.

Another restriction of the above model is that all relevant variables must be included in the model (causal sufficiency). Hoyer et al. (2008b) extended the above model by allowing for hidden variables. This leads to an overcomplete basis ICA model, meaning that there are more independent non-Gaussian sources than observed mixtures. While there exist methods for estimating overcomplete basis ICA models, those methods which achieve the required accuracy do not scale well. Additionally, the solution is again only unique up to ordering, scaling, and sign, and when including hidden variables the ordering-ambiguity cannot be resolved and in some cases leads to several observationally equivalent models, just as in the cyclic case above.

We note that it is also possible to combine the approach of section 2 with that described here. That is, if some of the shocks are Gaussian or close to Gaussian, it may be advantageous to use a combination of constraint-based search and non-Gaussianity-based search. Such an approach was proposed in Hoyer et al. (2008a). In particular, the proposed method does not make any assumptions on the distributions of the VAR-residuals $\mathbf{u}_t$. Basically, the PC algorithm (see Section 2) is run first, followed by utilization of whatever non-Gaussianity there is to further direct edges. Note that there is no need to know in advance which shocks are non-Gaussian since finding such shocks is part of the algorithm.

Finally, we need to point out that while the basic ICA-based approach does not require the faithfulness assumption, the extensions discussed at the end of this section do.

4. Nonparametric setting

4.1. Theory

Linear systems dominate VAR, SVAR, and more generally, multivariate time series models in econometrics. However, it is not always the case that we know how a variable X may cause another variable Y. It may be the case that we have little or no *a priori* knowledge

about the way how Y depends on X. In its most general form we want to know whether X is independent of Y conditional on the set of potential graphical parents Z, i.e.

$$H_0 : Y \perp\!\!\!\perp X \mid Z, \tag{17}$$

where Y, X, Z is a set of time series variables. Thereby, we do not per se require an *a priori* specification of how Y possibly depends on X. However, constraint based algorithms typically specify conditional independence in a very restrictive way. In continuous settings, they simply test for nonzero partial correlations, or in other words, for linear (in)dependencies. Hence, these algorithms will fail whenever the data generation process (DGP) includes nonlinear causal relations.

In search for a more general specification of conditional independency, Chlaß and Moneta (2010) suggest a procedure based on nonparametric density estimation. Therein, neither the type of dependency between Y and X, nor the probability distributions of the variables need to be specified. The procedure exploits the fact that if two random variables are independent of a third, one obtains their joint density by the product of the joint density of the first two, and the marginal density of the third. Hence, hypothesis test (17) translates into:

$$H_0 : \frac{f(Y,X,Z)}{f(XZ)} = \frac{f(YZ)}{f(Z)}. \tag{18}$$

If we define $h_1(\cdot) := f(Y,X,Z)f(Z)$, and $h_2(\cdot) := f(YZ)f(XZ)$, we have:

$$H_0 : h_1(\cdot) = h_2(\cdot). \tag{19}$$

We estimate h_1 and h_2 using a kernel smoothing approach (see Wand and Jones, 1995, ch.4). Kernel smoothing has the outstanding property that it is insensitive to autocorrelation phenomena and, therefore, immediately applicable to longitudinal or time series settings (Welsh et al., 2002).

In particular, we use a so-called *product kernel* estimator:

$$\begin{aligned}
\hat{h}_1(x,y,z;b) &= \frac{1}{N^2 b^{m+d}} \left\{ \sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) K\left(\frac{Y_i - y}{b}\right) K\left(\frac{Z_i - z}{b}\right) \right\} \left\{ \sum_{i=1}^{n} K_p\left(\frac{Z_i - z}{b}\right) \right\} \\
\hat{h}_2(x,y,z;b) &= \frac{1}{N^2 b^{m+d}} \left\{ \sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) K_Z\left(\frac{Z_i - z}{b}\right) \right\} \left\{ \sum_{i=1}^{n} K\left(\frac{Y_i - y}{b}\right) K_p\left(\frac{Z_i - z}{b}\right) \right\},
\end{aligned} \tag{20}$$

where X_i, Y_i, and Z_i are the i^{th} realization of the respective time series, K denotes the *kernel* function, b indicates a scalar bandwidth parameter, and K_p represents a product kernel[2].

So far, we have shown how we can estimate h_1 and h_2. To see whether these are different, we require some similarity measure between both conditional densities. There are different ways to measure the distance between a product of densities:

2. I.e. $K_p((Z_i - z)/b) = \prod_{j=1}^{d} K((Z_{j_i} - z_j)/b)$. For our simulations (see next section) we choose the kernel: $K(u) = (3 - u^2)\phi(u)/2$, with $\phi(u)$ the standard normal probability density function. We use a "rule-of-thumb" bandwidth: $b = n^{-1/8.5}$.

(i) The weighted Hellinger distance proposed by Su and White (2008):

$$d_H = \frac{1}{n}\sum_{i=1}^{n}\left\{1 - \sqrt{\frac{h_2(X_i, Y_i, Z_i)}{h_1(X_i, Y_i, Z_i)}}\right\}^2 a(X_i, Y_i, Z_i), \tag{21}$$

where $a(\cdot)$ is a nonnegative weighting function. Both the weighting function $a(\cdot)$, and the resulting test statistic are specified in Su and White (2008).

(ii) The Euclidean distance proposed by Szekely and Rizzo (2004) in their 'energy test':

$$d_E = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}||h_{1_i} - h_{2_j}|| - \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}||h_{1_i} - h_{1_j}|| - \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}||h_{2_i} - h_{2_j}||, \tag{22}$$

where $h_{1_i} = h_1(X_i, Y_i, Z_i)$, $h_{2_i} = h_2(X_i, Y_i, Z_i)$, and $||\cdot||$ is the Euclidean norm.[3]

Given these test statistics and their distributions, we compute the type-I error, or *p-value* of our test problem (19). If $Z = \emptyset$, the tests are available in R-packages `energy` and `cramer`. The Hellinger distance is not suitable here, since one can only test for $Z \neq \emptyset$.

For $Z \neq \emptyset$, our test problem (19) requires higher dimensional kernel density estimation. The more dimensions, i.e. the more elements in Z, the scarcer the data, and the greater the distance between two subsequent data points. This so-called *Curse of dimensionality* strongly reduces the accuracy of a nonparametric estimation (Yatchew, 1998). To circumvent this problem, we calculate the type-I errors for $Z \neq \emptyset$ by a local bootstrap procedure, as described in Su and White (2008, pp. 840-841) and Paparoditis and Politis (2000, pp. 144-145). Local bootstrap draws repeatedly with replacement from the sample and counts how many times the bootstrap statistic is larger than the test statistic of the entire sample. Details on the local bootstrap procedure ca be found in appendix A.

Now, let us see how this procedure fares in those time series settings, where other testing procedures failed - the case of nonlinear time series.

4.2. Simulation Design

Our simulation design should allow us to see how the search procedures of 4.1 perform in terms of *size* and *power*. To identify size properties (type-I error), H_0 (19) must hold everywhere. We call data generation processes for which H_0 holds everywhere, *size-DGPs*. We induce a system of time series $\{V_{1,t}, V_{2,t}, V_{3,t}\}_{t=1}^{n}$ whereby each time series follows an autoregressive process AR(1) with $a_1 = 0.5$ and error term $e_t \sim N(0, 1)$, for instance, $V_{1,t} = a_1 V_{1,t-1} + e_{V_1,t}$. These time series may cause each other as illustrated in Fig. 1.

Therein, $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,t-1}$, since $V_{1,t-1}$ d-separates $V_{1,t}$ from $V_{2,t}$, while $V_{2,t} \perp\!\!\!\perp V_{3,s}$, for any t and s. Hence, the set of variables Z, conditional on which two sets of variables X and

3. An alternative Euclidean distance is proposed by Baringhaus and Franz (2004) in their Cramer test. This distance turns out to be $d_E/2$. The only substantial difference from the distance proposed in (ii) lies in the method to obtain the critical values (see Baringhaus and Franz 2004).

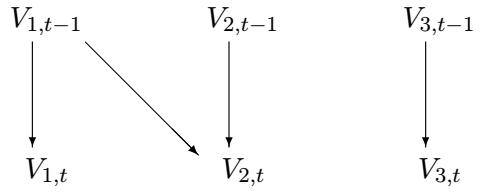

Figure 1: Time series DAG.

Y are independent of each other, contains zero elements, i.e. $V_{2,t} \perp\!\!\!\perp V_{3,t-1}$, contains one element, i.e. $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,t-1}$, or contains two elements, i.e. $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,t-1}, V_{3,t-1}$.

In our simulations, we vary two aspects. The first aspect is the *functional form of the causal dependency*. To systematically vary nonlinearity and its impact, we characterize the causal relation between, say, $V_{1,t-1}$ and $V_{2,t}$, in a polynomial form, i.e. via $V_{2,t} = f(V_{1,t-1}) + e$, where $f = \sum_{j=0}^{p} b_j V_{1,t-1}^j$. Herein, j reflects the degree of nonlinearity, while b_j would capture the impact nonlinearity exerts. For polynomials of any degree, we set only $b_p \neq 0$. An additive error term e completes the specification.

The second aspect is the number of variables in Z conditional on which X and Y can be independent. Either zero, one, but maximally two variables may form the set $Z = \{Z_1, \ldots, Z_d\}$ of conditioned variables; hence Z has cardinality $\#Z = \{0, 1, 2\}$.

To identify power properties, H_0 must not hold anywhere, i.e. $X \not\perp\!\!\!\perp Y|Z$. We call data generation processes where H_0 does not hold anywhere, *power-DGPs*. Such DGPs can be induced by (i) a direct path between X and Y which does not include Z, (ii) a common cause for X and Y which is not an element of Z, or (iii) a "collider" between X and Y belonging to Z.[4] As before, we vary the functional form f of these causal paths polynomially where again, only $b_p \neq 0$. Third, we investigate different cardinalities $\#Z = \{0, 1, 2\}$ of set Z conditional on which X and Y become dependent.

4.3. Simulation Results

Let us start with $\#Z = 0$, that is, $H_0 := X \perp\!\!\!\perp Y$. Table 2 reports our simulation results for both size and power DGPs. Rejection frequencies are reported for three different tests, for a theoretical level of significance of 0.05, and 0.1.

Take the first line of Table 2. For *size DGPs*, H_0 holds everywhere. A test performs accurately if it rejects H_0 in accordance with the respective theoretical significance level. We see that the energy test rejects H_0 slightly more often than it should ($0.065 > 0.05; 0.122 > 0.1$), whereas the Cramer test does not reject H_0 often enough ($0.000 < 0.05, 0.000 < 0.1$). In comparison to the standard parametric Fisher's z, we see that the latter rejects H_0 much more often than it should. The energy test keeps the type-I error most accurately. Contrary to both nonparametric tests, the parametric procedure leads us to suspect a lot more causal relationships than there actually are, if $\#Z = 0$.

How well do these tests perform if H_0 does not hold anywhere? That is, how accurately do they reject H_0 if it is false *(power-DGPs)*? For linear time series, we see that the nonparametric energy test has nearly as much power as Fisher's z. For nonlinear time

4. An example of collider is displayed in Figure 1: $V_{2,t}$ forms a collider between $V_{1,t-1}$ and $V_{2,t-1}$.

Table 2: Proportion of rejection of H_0 (no conditioned variables)

	Energy	Cramer	Fisher	Energy	Cramer	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S0.1 (ind. time series)	0.065	0.000	0.151	0.122	0.000	0.213
Power DGPs						
P0.1 (time series linear)	0.959	0.308	0.999	0.981	0.462	1
P0.2 (time series quadratic)	0.986	0.255	0.432	0.997	0.452	0.521
P0.3 (time series cubic)	1	0.905	1	1	0.975	1
P0.3 (time series quartic)	1	0.781	0.647	1	0.901	0.709

Note: length series (n) = 100; number of iterations = 1000

series, the energy test clearly outperforms Fisher's z[5]. As it did for *size*, Cramer's test generally underperforms in terms of power. Interestingly, its power appears to be higher for higher degrees of nonlinearity. In summary, if one wishes to test for marginal independence without any information on the type of a potential dependence, one would opt for the energy test. It has a size close to the theoretical significance level, and has power similar to a parametric specification.

Let us turn to $\#Z = 1$, where $H_0 := X \perp\!\!\!\perp Y|Z$, for which the results are shown in Table 3. Starting with *size DGPs*, tests based on Hellinger and Euclidian distance slightly underreject H_0 whereas for the highest polynomial degree, the Hellinger test strongly overrejects H_0. The parametric Fisher's z slightly overrejects H_0 in case of linearity, and for higher degrees, starts to underreject H_0.

Table 3: Proportion of rejection of H_0 (one conditioned variable)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S1.1 (time series linear)	0.035	0.035	0.062	0.090	0.060	0.103
S1.2 (time series quadratic)	0.040	0.020	0.048	0.065	0.035	0.104
S1.3 (time series cubic)	0.010	0.010	0.050	0.020	0.015	0.093
S1.4 (time series quartic)	0.13	0	0.023	0.2	0.1	0.054
Power DGPs						
P1.1 (time series linear)	0.875	0.910	0.999	0.925	0.950	1
P1.2 (time series quadratic)	0.905	0.895	0.416	0.940	0.950	0.504
P1.3 (time series cubic)	0.990	1	1	1	1	1
P1.4 (time series quartic)	0.84	0.995	0.618	0.91	0.995	0.679

Note: n = 100; number of iterations = 200; number of bootstrap iterations (I) = 200

Turning to *power DGPs*, Fisher's z suffers a dramatic loss in power for those polynomial degrees which depart most from linearity, i.e. quadratic, and quartic relations. Nonpara-

5. For cubic time series, Fisher's z performs as well as the energy test does. This may be due to the fact that a cubic relation resembles more to a line than other polynomial specifications do.

metric tests which do not require linearity have high power in absolute terms, and nearly twice as much as compared to Fisher's z. The power properties of the nonparametric procedures indicate that our local bootstrap succeeds in mitigating the Curse of dimensionality. In sum, nonparametric tests exhibit good power properties for $\#Z = 1$ whereas Fisher's z would fail to discover underlying quadratic or quartic relationships in some 60%, and 40% of the cases, respectively.

The results for $\#Z = 2$ are presented in Table 4. We find that both nonparametric tests have a size which is notably smaller than the theoretical significance level we induce. Hence, both have a strong tendency to underreject H_0. Turning to *power DGPs*, we find that the Euclidean test still has over 90% power to correctly reject H_0. For those polynomial degrees which depart most from linearity, i.e. quadratic and quartic, the Euclidean test has three times as much power as Fisher's z. However, the Hellinger test performs even worse than Fisher's z. Here, it may be the Curse of dimensionality which starts to show an impact.

Table 4: Proportion of rejection of H_0 (two conditioned variables)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	level of significance 5%			*level of significance* 10%		
Size DGPs						
S2.1 (time series linear)	0.006	0.020	0.050	0.033	0.046	0.102
S2.2 (time series quadratic)	0.000	0.010	0.035	0.000	0.040	0.087
S2.3 (time series cubic)	0	0.007	0.056	0	0.007	0.109
S2.4 (time series quartic)	0.006	0	0.031	0.013	0	0.067
Power DGPs						
P2.1 (time series linear)	0.28	0.92	1	0.4	0.973	1
P2.2 (time series quadratic)	0.170	0.960	0.338	0.250	0.980	0.411
P2.3 (time series cubic)	0.667	1	1	0.754	1	1
P2.4 (time series quartic)	0.086	0.946	0.597	0.133	0.966	0.665

Note: $n = 100$; number of iterations = 150; number of bootstrap iterations (I) = 100

To sum up, we can say that both marginal independencies, and higher dimensional conditional independencies, i.e. ($\#Z = 1, 2$) are best tested for using Euclidean tests. The Hellinger test seems to be more affected by the Curse of dimensionality. We see that our local bootstrap procedure mitigates the latter, but we admit that the number of variables our nonparametric procedure can deal with is very small. Here, it might be promising to opt for semiparametric (Chu and Glymour, 2008), rather than nonparametric procedures which combine parametric and nonparametric approaches.

5. Conclusions

The difficulty of learning causal relations from passive, that is non-experimental, observations is one of the central challenges of econometrics. Traditional solutions involve the distinction between structural and reduced form model. The former is meant to formalize the unobserved data generating process, whereas the latter aims to describe a simpler transformation of that process. The structural model is articulated hinging on *a priori*

economic theory. The reduced form model is formalized in such a way that it can be estimated directly from the data. In this paper, we have presented an approach to identify the structural model which minimizes the role of *a priori* economic theory and emphasizes the need of an appropriate and rich statistical model of the data. Graphical causal models, independent component analysis, and tests of conditional independence are the tools we propose for structural identification in vector autoregressive models. We conclude with an overview of some important issues which are left open in this domain.

1. *Specification of the statistical model.* Data driven procedures for SVAR identification depend upon the specification of the (reduced form) VAR model. Therefore, it is important to make sure that the estimated VAR model is an accurate description of the dynamics of the included variables (whereas the contemporaneous structure is intentionally left out, as seen in section 1.2). The usual criterion for accuracy is to check that the model estimates residuals conform to white noise processes (although serial independence of residuals is not a sufficient criterion for model validation). This implies stable dependencies captured by the relationships among the modeled variables, and an unsystematic noise. It may be the case, as in many empirical applications, that different VAR specifications pass the model checking tests equally well. For example, a VAR with Gaussian errors and p lags may fit the data equally well as a VAR with non-Gaussian errors and q lags and these two specifications justify two different causal search procedures. So far, we do not know how to adjudicate among alternative and seemingly equally accurate specifications.

2. *Background knowledge and assumptions.* Search algorithms are based on different assumptions, such as, for example, causal sufficiency, acyclicity, the Causal Markov Condition, Faithfulness, and/or the existence of independent components. Maybe, background knowledge could justify some of these assumptions and reject others. For example, institutional or theoretical knowledge about an economic process might inform us that Faithfulness is a plausible assumption in some contexts rather than in others, or instead, that one should expect feedback loops if data are collected at certain levels of temporal aggregation. Yet, if background information could inform us here, this might again provoke a problem of circularity mentioned at the outset of the paper.

3. *Search algorithms in nonparametric settings.* We have provided some information on which nonparametric test procedures might be more appropriate in certain circumstances. However, it is not clear which causal search algorithms are most efficient in exploiting the nonparametric conditional independence tests proposed in Section 4. The more variables the search algorithm needs to be informed about at the same point of the search, the higher the number of conditioned variables, and hence, the slower, or the more inaccurate, the test.

4. *Number of shocks and number of variables.* To conserve degrees of freedom, SVARs rarely model more than six to eight time series variables (Bernanke et al., 2005, p.388). It is an open question how the procedures for causal inference we reviewed can be applied to large scale systems such as dynamic factor models. (Forni et al., 2000)

5. *Simulations and empirical applications.* Graphical causal models for identifying SVARs, equivalent or similar to the search procedures described in section 2, have been applied to several sets of macroeconomic data (Swanson and Granger, 1997; Bessler and Lee, 2002; Demiralp and Hoover, 2003; Moneta, 2004; Demiralp et al., 2008; Moneta, 2008; Hoover et al., 2009). Demiralp and Hoover (2003) present Monte Carlo simulations to

evaluate the performance of the PC algorithm for such an identification. There are no simulation results so far about the performance of the alternative tests on residual partial correlations presented in section 2.2. Moneta et al. (2010) applied an independent component analysis as described in section 3, to microeconomic US data about firms' expenditures on R&D and performance, as well as to macroeconomic US data about monetary policy and its effects on the aggregate economy. Hyvärinen et al. (2010) assess the performance of independent component analysis for identifying SVAR models. It is yet to be established how independent component analysis applied to SVARs fares compared to graphical causal models (based on the appropriate conditional independence tests) in non-Gaussian settings. Nonparametric tests of conditional independence, as those proposed in section 4, have been applied to test for Granger non-causality (Su and White, 2008), but there are not yet any applications where these test results inform a graphical causal search algorithm. Overall, there is a need for more empirical applications of the procedures described in this paper. Such applications will be useful to test, compare, and improve different search procedures, to suggest new problems, and obtain new causal knowledge.

6. Appendix

6.1. Appendix 1 - Details of the bootstrap procedure from 4.1.

(1) Draw a bootstrap sampling Z_t^* (for $t = 1, \ldots, n$) from the estimated kernel density $\hat{f}(z) = n^{-1}b^{-d}\sum_{t=1}^{n} K_p((Z_t - z)/b)$.

(2) For $t = 1, \ldots, n$, given Z_t^*, draw X_t^* and Y_t^* *independently* from the estimated kernel density $\hat{f}(x|Z_t^*)$ and $\hat{f}(y|Z_t^*)$ respectively.

(3) Using X_t^*, Y_t^*, and Z_t^*, compute the bootstrap statistic S_n^* using one of the distances defined above.

(4) Repeat steps (1) and (2) I times to obtain I statistics $\{S_{ni}^*\}_{i=1}^{I}$.

(5) The p-value is then obtained by:

$$p \equiv \frac{\sum_{i=1}^{I} 1\{S_{ni}^* > S_n\}}{I},$$

where S_n is the statistic obtained from the original data using one of the distances defined above, and $1\{\cdot\}$ denotes an indicator function taking value one if the expression between brackets is true and zero otherwise.

References

E. Baek and W. Brock. A general test for nonlinear Granger causality: Bivariate model. *Discussin paper, Iowa State University and University of Wisconsin, Madison*, 1992.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

B. S. Bernanke. Alternative explanations of the money-income correlation. In *Carnegie-Rochester Conference Series on Public Policy*, volume 25, pages 49–99. Elsevier, 1986.

B.S. Bernanke, J. Boivin, and P. Eliasz. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics*, 120(1):387–422, 2005.

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

O. J. Blanchard and D. Quah. The dynamic effects of aggregate demand and supply disturbances. *The American Economic Review*, 79(4):655–673, 1989.

O. J. Blanchard and M. W. Watson. Are business cycles all alike? *The American business cycle: Continuity and change*, 25:123–182, 1986.

N. Chlaß and A. Moneta. Can Graphical Causal Inference Be Extended to Nonlinear Settings? *EPSA Epistemology and Methodology of Science*, pages 63–72, 2010.

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

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

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

S. Demiralp, K. D. Hoover, and D. J. Perez. A Bootstrap method for identifying and evaluating a structural vector autoregression. *Oxford Bulletin of Economics and Statistics, 65, 745-767*, 2008.

M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

J. Faust and E. M. Leeper. When do long-run identifying restrictions give reliable results? *Journal of Business & Economic Statistics*, 15(3):345–353, 1997.

J. P. Florens and M. Mouchart. A note on noncausality. *Econometrica*, 50(3):583–591, 1982.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554, 2000.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

C. W. J. Granger. Testing for causality:: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:1–115, 1944.

C. Hiemstra and J. D. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance*, 49(5):1639–1664, 1994.

W. C. Hood and T. C. Koopmans. *Studies in econometric method, Cowles Commission Monograph, No. 14*. New York: John Wiley & Sons, 1953.

K. D. Hoover. *Causality in macroeconomics*. Cambridge University Press, 2001.

K. D. Hoover. The methodology of econometrics. *New Palgrave Handbook of Econometrics*, 1:61–87, 2006.

K. D. Hoover. Causality in economics and econometrics. In *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan, 2008.

K.D. Hoover, S. Demiralp, and S.J. Perez. Empirical Identification of the Vector Autoregression: The Causes and Effects of US M2. In *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, pages 37–58. Oxford University Press, 2009.

P. O. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008a.

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

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

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a Structural Vector Autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11: 1709–1731, 2010.

S. Johansen. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control*, 12:231–254, 1988.

S. Johansen. Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

S. Johansen. Cointegration: An Overview. In *Palgrave Handbook of Econometrics. Volume 1. Econometric Theory*, pages 540–577. Palgrave Macmillan, 2006.

R. G. King, C. I. Plosser, J. H. Stock, and M. W. Watson. Stochastic trends and economic fluctuations. *American Economic Review*, 81:819–840, 1991.

T. C. Koopmans. *Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph, No. 10*. New York: John Wiley & Sons, 1950.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by Independent Components Analysis. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence (UAI-2008)*, Helsinki, Finland, 2008.

R. E. Lucas. Econometric policy evaluation: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, volume 1, pages 19–46. Elsevier, 1976.

H. Lütkepohl. Vector Autoregressive Models. In *Palgrave Handbook of Econometrics. Volume 1. Econometric Theory*, pages 477–510. Palgrave Macmillan, 2006.

A. Moneta. Graphical Models for Structural Vector Autoregressions. *LEM Papers Series, Sant'Anna School of Advanced Studies, Pisa*, 2003.

A. Moneta. Identification of monetary policy shocks: a graphical causal approach. *Notas Económicas, 20, 39-62*, 2004.

A. Moneta. Graphical causal models and VARs: an empirical assessment of the real business cycles hypothesis. *Empirical Economics*, 35(2):275–300, 2008.

A. Moneta, D. Entner, P.O. Hoyer, and A. Coad. Causal inference by independent component analysis with applications to micro-and macroeconomic data. *Jena Economic Research Papers*, 2010:031, 2010.

E. Paparoditis and D. N. Politis. The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, 52(1):139–159, 2000.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, Cambridge, 2000.

M. Reale and G. T. Wilson. Identification of vector AR models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications, 10, 49-65*, 2001.

T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In *Computation, causation and discovery*. AAAI Press and MIT Press, Menlo Park, 1999.

R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

M. D. Shapiro and M. W. Watson. Sources of business cycle fluctuations. *NBER Macroeconomics annual*, 3:111–148, 1988.

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

S. Shimizu, A. Hyvärinen, Y. Kawahara, and T. Washio. A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

C. A. Sims. Macroeconomics and Reality. *Econometrica, 48, 1-47*, 1980.

C. A. Sims. An autoregressive index model for the u.s. 1948-1975. In J. Kmenta and J.B. Ramsey, editors, *Large-scale macro-econometric models: theory and practice*, pages 283–327. North-Holland, 1981.

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

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, 2008.

P. Suppes. A probabilistic theory of causation. *Acta Philosophica Fennica, XXIV*, 1970.

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

G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

M. P. Wand and M. C. Jones. Kernel smoothing. *Chapman&Hall Ltd., London*, 1995.

A. H. Welsh, X. Lin, and R. J. Carroll. Marginal Longitudinal Nonparametric Regression. *Journal of the American Statistical Association*, 97(458):482–493, 2002.

H. White and X. Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193, 2010.

N. Wiener. The theory of prediction. *Modern mathematics for engineers, Series*, 1:125–139, 1956.

S. Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998.