No. 2007-11
August 27, 2007
Economics
The Open Access, Open Assessment E-Journal

Learning Causal Relations in Multivariate Time Series Data

Pu Chen and Hsiao Chihying

Bielefeld University, Germany

Abstract:

Applying a probabilistic causal approach, we define a class of time series causal models (TSCM) based on stationary Bayesian networks. A TSCM can be seen as a structural VAR identified by the causal relations among the variables. We classify TSCMs into observationally equivalent classes by providing a necessary and sufficient condition for the observational equivalence. Applying an automated learning algorithm, we are able to consistently identify the data-generating causal structure up to the class of observational equivalence. In this way we can characterize the empirical testable causal orders among variables based on their observed time series data. It is shown that while an unconstrained VAR model does not imply any causal orders in the variables, a TSCM that contains some empirically testable causal orders implies a restricted SVAR model. We also discuss the relation between the probabilistic causal concept presented in TSCMs and the concept of Granger causality. It is demonstrated in an application example that this methodology can be used to construct structural equations with causal interpretations.

JEL: C1
Keywords: Automated Learning, Bayesian Network, Inferred Causation, VAR, Wage-Price Spiral

Correspondence:
Pu Chen, Faculty of Economics, Bielefeld University, PO Box 10 01 31, 33501 Bielefeld, Germany, email: pchen@wiwi.uni-bielefeld.de,Tel.: 49 521 106 4875

Contents

1 Introduction

Since the development of the successful learning algorithms for Bayesian networks, the probabilistic causal approach attracts more and more attention of the scientific community[1]. Spirtes, Glymour, and Scheines (2001) provide a detailed description of learning Bayesian networks through sequential tests and the causal interpretation of the test results. Pearl (2000) gives a rigorous account of the probabilistic approach to causality. Heckerman, Geiger, and Chickering (1995) provide the Bayesianian technique for learning Bayesian networks from data. Despite the controversial debate on this Bayesian network causal approach[2], the automated causal inference based on Bayesian network models becomes an effective instrument to assess causal relations empirically.

Recently, these graphical models have found their way into the literature on time series analysis and econometrics. Dahlhaus (2000) gives a graphical interpretation of the conditional independence among the elements of multivariate time series. Bach and Jordan (2004) present graphical models for multivariate time series in the frequency domain. Eichler (2003) gives a graphical presentation of the Granger causality among the elements of multivariate time series. Some pioneering works of graphical models in econometrics can be found in Glymour and Spirtes (1988). Hoover (2005) sketches the application of the Bayesian network technique for identifying structural VAR models. Swanson and Granger (1997) apply a similar concept to identity the causal chain in VAR residuals. Demiralp and Hoover (2004) apply the Bayesian network method to VAR residuals to infer the causal order in the money demand and the monetary transmission mechanism.

Following this line of research, in this paper we develop a causal model for multivariate time series data. We apply the probabilistic causal approach to define causal models for multivariate time series. Under reasonable assumptions on the causal structures for time series, TSCMs become statistically assessable. Further we show that these TSCMs are equivalent to SVAR models. In this way, we give a causal theoretical justification for the application of the automatic inference to identify a SVAR as described in Hoover (2005). We interpret a SVAR in terms of a contemporaneous causal structure and a temporal causal structure. A two-step procedure is developed to learn the contemporaneous and the temporal causal structure of a multivariate time series causal model.

The rest of the paper is organized as follows.

[1] Although inferring causal relations used to be the primary target of statistical analysis, this ambition was abandoned for a long time. See Pearl (2000) for more details.

[2] see Cartwright (2001) and Pearl (2000) p. 41 for more details.

In Section 2 we review the basic idea of the inferred causation. Here we focus on the causal interpretation of the Bayesian network models and their relations to linear recursive structural models. Within the class of linear recursive structural models we discuss in detail the structure of inferrable causation and the model equivalence. In Section 3 we extend the concept of the causal models to time series data and define time series causal models. Here we show the equivalence of TSCMs and SVAR models, and discuss the relation between the Granger causality and the probabilistic causal dependence. In Section 4 we present a two step procedure to estimate the time series causal models from the observed time series data. We show the consistence of the procedure and document some simulation results to assess the small sample properties of the procedure, as well as the effectiveness of the procedure in recovering the true causal order. Section 6 is devoted to an illustrative application example of the TSCM. The last section concludes.

2 Inferred Causation

2.1 A Model Selection Approach to Inferred Causation

A fundamental assumption of the method of inferred causation is that, as given in Definition 2 in Pearl and Verma (1991): the casual relations among a set of variables U can be modelled in a directed acyclic graph(DAG) D and a set of parameters Θ_D, compatible with D. Θ_D assigns a function $x_i = f_i(pa(x_i), \epsilon_i)$ and a probability measure g_i to each $x_i \in U$, where $pa(x_i)$ are parents of x_i in D and each $epsilon_i$ is a random disturbance distributed according to g_i independently of the other ϵ's and of any preceding x_j: $0 < j < i$.

The probability measure compatible with D is called to satisfy the Markov condition in Pearl (2000) p.16. The Markov condition implies in particular that the disturbance ϵ_i are independent form other ϵ's. In addition to the Markov condition, the minimality of the causal structure[3], D, and the stability of the distribution are two key assumptions on the data-generating causal model to rule out the ambiguity of the statistical inference in recovering the data-generating causal model.[4]. Further, a DAG with a Probability measure P that satisfy the Markov condition with respect to the DAG(See Fig.1 for examples.) prescribes an ordering of the variables in the DAG and the factorization of the joint distribution of the variables as the product of

[3]See Definition 5 in Pearl and Verma (1991)

[4]It is still an ongoing debate whether causality can be formulated in such assumptions. See Cartwright (2001), Pearl (2000), Spirtes et al. (2001) Freedman and Humphreys (1998) for more discussion. Spirtes et al. (2001) took an axiomatic approach to pave the logical basis for the method of inferred causation.

the conditional distributions. A sparse DAG implies in particular a set of conditional dependence and independence among variables. In (a) and (b) of Fig.1 A and C is said to be d-separated by B. This implies that for all compatible distributions with the DAGs A and C would be dependent, but conditioning on B, they would be independent. In this case B is said to screen A from B. In (c) of Fig.1 A and C is not d-separated by B. This implies that at least for one distribution compatible with the DAG, A and C would be independent, but conditioning on B they would become dependent[5]. B is the effect of A and C. Here B is called an unshielded collider on the path ABC. In the literature an unshield collider is also called a v structure, because it consists of two converging arrows whose ends are not connected. A shielded collider would have a direct link between A and C.

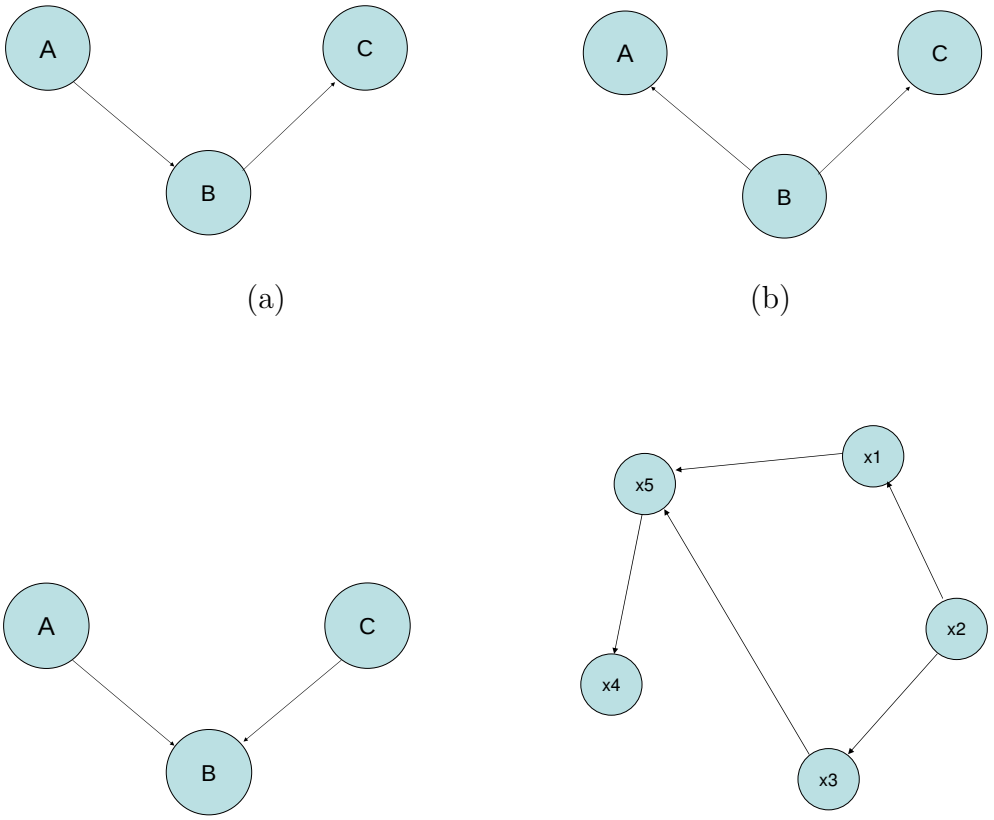

Figure 1: Influence Diagram

Under the Markov assumption, a compatible distribution of a DAG can be factorized into the conditional distributions according to the DAG. Hence we

[5]See Pearl (2000) p.18.

know that the DAG (d) in Fig.1 implies that the joint distribution can be calculated as follows.

$$\begin{aligned} & f(x_{1t}, x_{2t}, x_{3t}, x_{4t}, x_{5t}) \\ = \; & f(x_{4t}|x_{5t}) f(x_{5t}|x_{1t}, x_{3t}) f(x_{3t}|x_{2t}) f(x_{1t}|x_{2t}) f(x_{2t}), \end{aligned} \tag{2.1}$$

This implies the following conditional independence: given x_{5t}, x_{4t} is independent on other variables; given x_{1t} and x_{3t}, x_{5t} is independent on x_{2t}; and given x_{2t}, x_{3t} is independent on x_{1t}.

The fundamental assumption of the method of inferred causation translates the problem to infer causal relations among variables into a statistical problem to recover the true data generating DAG model using the observed data, and then to interpret the directed edges in the DAG as causal relations.

The implication of a DAG on the patterns of the conditional dependence and independence invites inference of the data generating DAG from these patterns of the conditional dependence and independence. Identifying the underlying DAG from the patterns of conditional independence and dependence has been the main research activity in the area of inferred causation. We will give a more detailed description about it in the next section.

Alternatively, consistent model selection criteria can also be used to identify the data generating DAG, if the data generating DAG is under the set of models to be selected. The assumption that the data generating DAG is under the set of DAG models under consideration is called the causal sufficiency assumption[6].

Therefore, under causal sufficiency applying a consistent model selection criterion to search over all possible DAG models will identify the data generating DAG or its observationally equivalent models consistently.

In this paper we will use this method to uncover the data generating DAG. The statistical process of uncovering the data generating DAG is called learning of DAG in the literature.

In example (d) in Fig.1 we will search over all DAG models consisting of the five variables $x_{5t}, x_{4t}, x_{3t}, x_{2t}, x_{1t}$. A consistent model selection criterion evaluates a model by the sum of its likelihood and a penalty on the dimensionality of the model. The likelihood is the leading term in this sum such that all misspecified models will not be selected asymptotically and the penalty term

[6]If some variables are not observed(these kind of variables are called latent variables), then the data generating DAG may not be within the set of DAG models to be investigated. The method of inferred causation can be used to detect the existence of latent variables. We will not discuss this issue in this paper. We consider here only the cases under causal sufficiency assumption.

will go to infinite as $T \to \infty$, such that the probability to select a model with too many parameters will converge to zero.

In this context, statistically learning of the causal order is equivalent to searching for the most parsimonious model that can account for the joint distribution of the variables in the class of all possible recursive models.

Now it is of interest to ask:

- If data are generated from a causal model, can statistical procedures always uniquely identify this causal model?
- If a causal model cannot be uniquely identified by statistical procedures, which causal properties of the causal model can be identified by statistical procedures?
- How effective is a statistical learning procedure?

The answers to these questions are the main research issues of the probabilistic causal approach. The first and the second question concern the observational equivalence of causal models and the assumptions of causal models. The third one concerns the efficiency of algorithms to learn causal relations implied in the observed data. Pearl (2000), Spirtes et al. (2001) and Heckerman et al. (1995) provide the most detailed and up-to-date accounts in this area.

Observationally equivalent models will generate data with identical statistical properties. Therefore, statistical method can only identify the underlying DAGs up to the observationally equivalent classes. For the observational equivalence we quote the results in Pearl (2000) p.19.

Proposition 2.1 *[Observational Equivalence] Two DAGs(models) are observationally equivalent if and only if they have the same skeleton and the same set of v-structures, that is two converging arrows whose tails are not connected by an arrow (Verma and Pearl 1990).*

Since statistical method cannot differ the observationally equivalent models from each other from the data, not every causal direction in a DAG can always be identified according to this Proposition. Only those causal directions in a DAG can be identified, if they constitute v structures or if their change would result in new v-structures or cycles. Consequently, if a data generating DAG has observationally equivalent models, i.e. there exists some arrows in the DAG, the change of whose directions will not lead to a new v structure or cycles, the direction of these arrows in the DAG cannot be uniquely inferred from the data. The existence of observational equivalence places a

limit on the ability of statistical method to identify the the directionality of dependence.

Given a set of data generated from a causal model, a statistical procedure can principally identify all the conditional independence. However, the statistical procedure cannot differ whether this kind of independence is due to a lack of the edge in the DAG of the causal model or due to particularly chosen parameter values of the DAS such that the edge in this case implies the independence. To rule out this ambiguity, Pearl (2000) assumes that all the identified conditional independence are due to lack of edges in the DAG of the causal model. This assumption is called stability condition in Pearl (2000). In Spirtes et al. (2001) it is called faithfulness condition. This assumption is therefore important for interpreting the conditional dependence and independence as causal relations.

2.2 DAGs and Structural Models

It can be generally shown that if an n-vector X is jointly normally distributed, a DAG model of X is equivalent to a linear recursive simultaneous equation model (SEM).

$$x_j = \sum_{k=1}^{j-1} a_{jk} x_k + \epsilon_j, \qquad \text{j = 1,2,...n} \tag{2.2}$$

where ϵ_j is independently normally distributed. We call (2.2) a linear causal model. We put this fact in the following proposition[7].

Proposition 2.2 *If a set of variables X are jointly normal $X \sim N(0, \Omega)$, a DAG model for X can be equivalently formulated as a linear recursive simultaneous equations model that is represented by a lower triangular coefficient matrix A with 1s along the principle diagonal. Any nonzero elements in this coefficient matrix, say a_{jk} correspond to a directed edge form variable k to variable j.*

$$A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -a_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -a_{n1} & -a_{n2} & \dots & 1 \end{pmatrix}, \tag{2.3}$$

where A is the inversion of the triangular decomposition matrix of Ω with $A\Sigma A' = D$ and D is a diagonal matrix.

[7]Bayesian network models can be used to encode any joint distributions. Therefore, they can also be applied to nonlinear models. Because linear models are often used in econometrics we discuss here only linear models.

Proof: Let Ω be the covariance matrix of X. A Bayesian network model for X is a factorization of the joint distribution as product of the conditional distributions of the components of X in a given order. Because conditional distributions of jointly normal distributed random variables are normal and the conditional means are linear functions of conditioning variables, a Bayesian network model for jointly normal distributed variables corresponds to a linear recursive simultaneous equations model. □

Remark 1 It is worth noting that using the rule given in Proposition 2.2 we can always get a unique corresponding DAG from a linear recursive simultaneous equations model. But from a DAG of jointly normally distributed variables we may sometimes get different linear recursive simultaneous equation models. For example the DAG of (c) in Fig. 1 can be written as:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -a_{ca} & -a_{cb} \end{pmatrix} \begin{pmatrix} X_a \\ X_b \\ X_c \end{pmatrix} = \begin{pmatrix} \epsilon_a \\ \epsilon_b \\ \epsilon_c \end{pmatrix} \tag{2.4}$$

or

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -a_{cb} & -a_{ca} \end{pmatrix} \begin{pmatrix} X_b \\ X_a \\ X_c \end{pmatrix} = \begin{pmatrix} \epsilon_b \\ \epsilon_a \\ \epsilon_c \end{pmatrix} \tag{2.5}$$

Such linear causal models presented by their coefficient matrices are trivially equivalent, because they present the same causal information and they differ only in the causally irrelevant order of their components. We call such linear causal models that correspond to the same DAG "trivially equivalent models".

Remark 2 Given the correspondence between a recursive SEM and a DAG, the parameter a_{ij} of the SEM corresponds to the edge from the vertex x_j to the vertex x_i. $a_{ij} = 0$ corresponds to the absence of the edge from the vertex x_j to the vertex x_i, which implies that x_j and x_i are conditionally independent, given the predecessors of x_i. Therefore, the more null restrictions a recursive model has, the simpler the corresponding DAG will be. Searching for the DAG with minimal structure is equivalent to searching for the most parsimonious recursive SEM for the data.

A useful property of multivariate normal distribution is that the conditional covariance, the conditional variance and the conditional correlation coefficient: $\sigma_{X_j,X_i|z}$, $\sigma_{X_j|z}$ and $\rho_{X_j,X_i|z}$ are all independent of the value z. Moreover the partial correlation coefficient is zero if and only if $(X_i \perp X_j|z)$[8]. Because we can estimate a recursive SEM by OLS, we have an important relation

[8] $(X_i \perp X_j|z)$ denotes that conditioned on z, X_i and X_j are independent.

between the parameter of the recursive SEM and the partial correlation coefficient(See also Pearl (2000) Chapter 2.):

$$r_{YX.Z} = \rho_{YX|Z}\frac{\sigma_{Y|Z}}{\sigma_{X|Z}}, \tag{2.6}$$

where $r_{YX.Z}$ is the regression coefficient of Y in the linear regression on X and Z

$$Y = aX + b_1 z_1 + b_2 z_2 + ... + b_k z_k \tag{2.7}$$

This means the coefficient a is given by $a = r_{YX.Z}$. This relation is very useful for deriving the results later.

2.3 Observational Equivalence and Inferrable Causation in SEMs

From Proposition 2.1 we know some arrows in a data generating DAG may not be identified due to the existence of observationally equivalent models. In this subsection we study how the condition of the observational equivalence is expressed by the parameters in linear causal models. A linear causal model is a recursive structural equation model, the upper triangular elements of the coefficient matrix are zeros (See Eq. (2.3)). A linear causal model is characterized through the zero restrictions on the parameters in the the lower triangular part of the coefficient matrix. Hence, when talk about zero restrictions, we mean the zero elements in the lower triangular part of the coefficient matrix.

If two different causal models can generate data with the same statistical property we will have problems to differentiate these two causal models by using statistical methods. Therefore, we have the following definition.

Definition 2.3 (Observationally Equivalent Causal Models) *If two different linear causal models can always generate identical joint distribution, they are called observationally equivalent.*

To the relation between two trivially observationally equivalent causal models we have the following proposition:

Proposition 2.4 (Interchange Rule 1) *Let A be a lower triangular (recursive) coefficient matrix of a linear causal model, and row i and row j be two adjacent rows of A with $j = i+1$. Let $A_{i\leftrightarrow j}$ be the lower triangular coefficient matrix obtained by interchange of i-th row and i-th column with j-th row and j-th column. If $a_{j,i} = 0$. then A and $A_{i\leftrightarrow j}$ are trivially observationally equivalent.*

Proof: Because the DAG of A and the DAG of $A_{i\leftrightarrow j}$ are identical, they are trivially observationally equivalent. □

Remark This interchange rule can be extended to the case of interchange of some consecutive rows and columns. The consecutive rows and columns are called block. Let A be a lower triangular (recursive) coefficient matrix of a linear causal model and i is the index of a block of rows and j is the index of the next block of rows with $j = i + 1$. Let $A_{i\leftrightarrow j}$ be the lower triangular coefficient matrix obtained by interchange of i-th block of rows and i-th block of columns with j-th block of rows and j-th block of columns. If $A_{j,i} = 0$, then A and $A_{i\leftrightarrow j}$ are trivially observationally equivalent, where $A_{j,i}$ is the submatrix in A consisting of j-th block of rows and i-th block of columns.

Proposition 2.5 (Structure of Nontrivial Observational Equivalence) *A linear causal model has nontrivial observational equivalence if and only if there are two adjacent rows whose zero elements are in the same columns or they have no zero elements.*

Proof: See Appendix.

Corollary 2.6 (Interchange Rule 2) *Let A be a lower triangular (recursive) coefficient matrix of a linear causal model. If there are two adjacent rows, i-th and j-th rows, whose zero elements are in the same columns, then the interchange of these two rows and the corresponding columns $A_{i\leftrightarrow j}$ constitute a new observationally equivalent causal model.*

Proof: See the proof of Proposition 2.5.

Remark 1 An interchange of two adjacent rows and columns implies that the order of the recursion between these two variables changes. It does not mean that the parameter values remain the same after change. They are freely varying parameters before and after the change. In terms of a graph, the interchange of two adjacent rows changes the direction of the edge between the two variables if they are connected by an edge.

Remark 2 Let A be a lower triangular (recursive) coefficient matrix of a linear causal model. If there is no zero restriction in the first block of i-rows ($i = 2, 3, 4, ...n$), then according to Corollary 2.6, any order of these i-rows will constitute an observationally equivalent model to A. Especially, when there are no zero restrictions in A at all, any permutation of the order of the elements of X constitutes an observationally equivalent model to A. In this case the order of the recursion does not provide any information about the causal direction.

Remark 3 As we know from Proposition 2.1, the existence of an observationally equivalent model can be characterized by v-structures. In terms of graphs Corollary 2.6 says that we can alter the direction of the arrow $x_i \to x_j$ to get an observationally equivalent model if x_i's parents are the parents of x_j. That x_i's parents are the parents of x_j implies the arrow $x_i \to x_j$ does not constitute a v structure, because all the tails of arrows into x_j are connected with x_i. For the same reason an arrow $x_j \to x_i$ does not constitute a v structure. Therefore the change the direction of the arrow $x_i \to x_j$ will not lead to a new v structure. Further, since all parents of x_i are parents of x_j, there is no path from x_i to x_j. The change in direction of the arrow $x_i \to x_j$ will not lead to a cycle. Therefore, the change of the direction of the arrow $x_i \to x_j$ generates an observationally equivalent model, since the change of the direction of the arrow $x_i \to x_j$ result in a DAG with the same skeleton and v-structures.

Following Remark 3 above, only the directions of edges whose change will lead to the change of the v-structures or lead to a cycle can be used to infer causal dependence. Other direction of edges in DAGs do not have any causal implication.

Corollary 2.7 (Observational Differentiability) *A linear causal model is called observationally distinguishable, if and only if there are no two adjacent rows, obtained through interchange rule 1, such that their zero elements are on the same columns or they have no zero elements.*

Remark Expressed in terms of DAG this corollary means simply that a causal model is observationally differentiable if and only if the DAG consists only of v-structures and those edges whose change will change the v-structures or will lead to a cycle.

Proposition 2.8 (Structure of Observational Equivalence) *An observationally indistinguishable linear causal model consists of one or more blocks (consecutive adjacent rows) in which some zero elements are on the same columns.*

Proof: Applying the result of the Proposition 2.5, we know that if a linear causal model is observationally indistinguishable, it must have one or more blocks in which zero elements are on the same columns. Applying interchange rule 2 any reordering within such blocks will generate a new observationally equivalent linear causal model. □

Because the causal ordering within such blocks are statistically not inferrable but the causal ordering between different blocks are statistically inferrable, we call these blocks simultaneous causal blocks. Based on this observation we can characterize the inferrable causal structure as follows.

Proposition 2.9 (Inferrable Causal Structure)

Assuming that observed data are generated by an unknown linear causal model, we have the following results:

- *If the data generating linear causal model is observationally distinguishable, the causal order of the variables can be inferred uniquely.*
- *If the data generating linear causal model has no zero restrictions in the recursive coefficient matrix, then there is only one simultaneous causal block. No causal order can be inferred from this model.*
- *If the data-generating linear causal model has zero restrictions and observationally equivalent models, then the causal order of the variables can be inferred up to the simultaneous causal blocks, and those causal directions whose change will alter the v-structures or will lead to a cycle can be inferred uniquely.*

Proof: Because linear causal models are recursive simultaneous equation models, their parameters can be consistently estimated by OLS[9]. Now the data generating causal model is a member of the set of all recursive simultaneous equations models. If it is observationally distinguishable, it can be identified consistently, by using a consistent model selection criterion over the set of all recursive simultaneous equations models.

If there are no restrictions on the data-generating linear causal model, following the Corollary 2.6 Remark 2, no causal order can be inferred from the data.

If a linear causal model has zero restrictions and observationally equivalent models, using a consistent model selection criterion, we can identify the observationally equivalent class of the data-generating causal model. Following Corollary 2.6 and the Proposition 2.8, the v-structures and the order of the simultaneous causal blocks can be consistently identified. □

3 Learning Bayesian Networks

As stated in Section 1, inferring causal relations on a set of variables is to uncover the underlying data generating DAG from the observed data of the variables.

Principally, we could evaluate every possible recursive model and find the one with the maximal criterion value. This is, however, only practicable if the

[9]See Dhrymes (1993) for details

number of variables is very small, because the number of all possible causal models grows explosively with the increase of the number of variables. For a system of 6 variables there are 3781503 possible models. Even the most powerful computers will reach their limit of computation with the increase of the number of variables in the system.

To solve this problem many heuristic algorithms are developed. There are now basically three kinds of solutions to this problem. One is based on sequential tests of partial correlation coefficients. The tests run from the lower order partial correlation coefficients in unconstrained models to the higher order partial correlation coefficients[10]. A limited version of this algorithm can be found in Swanson and Granger (1997). Hoover (2005) gives a very intuitive description of this procedure. Spirtes et al. (2001) provide a detailed discussion about these kinds of algorithms. Pearl (2000) presents a version of this algorithm, called the IC algorithm, as follows.(We quote Pearl (2000) p. 50)

IC Algorithm (Inductive Causation)

Input: P a stable[11] distribution on a set X of variables.
Output: a pattern (DAG) compatible with P.

- for each pair of variables $(X_i,X_j) \in X$, search a set S_{ij} such that $(X_i \perp X_j | S_{ij})$ holds in P. Construct an undirected graph G such that vertices X_i and X_j are connected with an edge if and only if no such set S_{ij} can be found.

- For each pair of nonadjacent variables X_i and X_j with a common neighbor X_k, check if $X_k \in S_{ij}$.
  If it is, then continue. If is is not, then add arrowheads pointing as X_k: $(X_i -> X_k < -X_j)$.

- In the partially directed graph that results, orient as many of the undirected edges as possible subject to two conditions: (i) the orientation should not create a new v structure; and (ii) the orientation should not create a directed cycle.

Remarks: Principally, the construction of DAG using this class of procedures is based on statistical tests. Therefore the probability to choose wrong

[10]See http://www.phil.cmu.edu/projects/tetrad/ for more details and software for this algorithm.

[11]Stability of a distribution means that the freely varying parameters of the data-generating causal models will assume parameter values other than zero (or the probability to assume the value zero is zero) in order that all identified zero parameter-values are interpreted as zero restrictions on the parameters. It is also known as faithfulness condition.

models equals the probability of type I errors of the test. However, since the tests are consistent, this procedure will consistently identify the true DAG, if the significance level of the tests converges to zero as the number of observations goes to infinite.

The second solution is based on the Bayesian approach of model averaging. Heckerman (1995) documents the basic technique of this approach. This technique combines the subjective knowledge with the information of the observed data to infer the causal relation among variables. These kinds of algorithms differ in the choice of criteria for the goodness of fit that is often called the score of a network, and in the choice of search strategy. Because the search problem is NP-hard[12] heuristic search algorithms such as greedy search, greedy search with restarts, best-fit search, and Monte-Carlo method are used[13]. The third solution uses classic model selection approach. Its implementation is similar to the Bayesian approach but without any use of a priori information. A network is evaluated according to information criteria such AIC and BIC. The search algorithms are similar to those in the Bayesian approach, such as greedy search, and greedy search with restart.

Greedy Search Algorithm:

Input: P a stable distribution on a set of variables X.
Output: a (DAG) compatible with P.

- Step 1 Start with a Bayesian network A_o.
- Step 2 Calculate the network score according to BIC/AIC/likelihood criterion.
- Step 3 Generate the local neighbour networks by either adding, removing or reversing an edge of the network A_o.
- Step 4 Calculate the scores for the local neighbour networks. Choose the one with the highest score as A_n. If the highest score is larger than that of A_o, go to Step 2 and update A_o by A_n. If the highest scores is less than that of the original A_o, output A_o.

Applying consistent model selection criterion in the greedy search algorithm implies that if the data generating linear causal model is statistically distinguishable, and the greedy search can find the global maximum, then it will uniquely identify the causal order. If the data generating causal model is not statistically distinguishable, the greedy search algorithm will uniquely identify the causal order among the simultaneous causal blocks and the v-structures consistently.

[12]See Heckerman (1995) for details.

[13]See Heckerman (1995) for details. A R-package "deal" for learning the Bayesian network using the Bayesian approach can be found at http://www.r-project.org/gR/

4 Time Series Causal Models

4.1 Extending the Linear Causal Models to Time Series Data

As we know, an n-dimensional multivariate time series can be generally represented by a sequence of random n-vector $\{X_t\}$ with a discrete index set $t \in I$ and each X_t has n elements indexed by $i \in \{1, 2, ..., n\}$. A linear causal model for the sequence $\{X_t\}$ will be a recursive model of $\{X_t\}$ in all its elements (indexed by t and i). In terms of graphs each vertex of the corresponding DAG represents a random element of X_{it}. Since we have only one observation for each random element X_{it}, many restrictions have to be imposed on this recursive model to make statistical inference possible. The task is now to formulate reasonable restrictions on the recursive model such that the resulting class of models are general enough to encompass most practically useful time series models and restrictive enough to allow statistical assessment. One naturally obvious restriction is the temporal causal constraint, i.e. the variable X_t cannot be a cause for $X_{t-\tau}$ for $\tau > 0$. This implies that the direction of an edge between two vertices in a DAG of time series causal models always goes from the vertex with an earlier time index to the vertex with a later time index or to a vertex with the same time index, but never the other way around. The time index provides here a natural causal direction here. Hence, a time series causal model can be formulated in the following way[14]:

$$\begin{pmatrix} A_{01} & 0 & \dots & 0 \\ A_{21} & A_{02} & & 0 \\ \vdots & & \ddots & \vdots \\ A_{T1} & A_{T2} & \dots & A_{0T} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}, \tag{4.8}$$

where $\epsilon_t, t = 1, 2, ..., T$ are vectors of independent random variables[15]. Obviously the temporal causal constraints are not enough, because there are still more unknown parameters in the coefficient matrix than the number of observations. A further reasonable constraint is the time invariance of the causal relation. This means the causal relation between the variables X_t and $X_{t-\tau}$ should be the same as the causal relation between X_{t+s} and $X_{t-\tau+s}$. This constraint implies that up to the initial conditions, parameters in each row of the coefficient matrix in (4.8) is the same, because they represent the causal relation between the variable of the current and the past variables.

[14] We have an explicit formulation of the initial conditions of the time series model: the model starts by $t = 1$ for simplicity of presentation.

[15] In the model above we have assumed that the random process have started at $t = 1$.

As $T \to \infty$, the equation (4.8) becomes a matrix equation with infinite dimension. The time invariance of the causal relation requires that each n rows of the coefficient matrix in (4.9) is the same, if it is read from the diagonal to the left.

$$\begin{pmatrix} \dots & A_2 & A_1 & A_0 & 0 & \dots & \dots & 0 \\ & \dots & A_2 & A_1 & A_0 & 0 & \dots & 0 \\ & & & & & \ddots & \ddots & \vdots \\ & & & \dots & A_2 & A_1 & A_0 & 0 \\ & & & & \dots & A_2 & A_1 & A_0 \end{pmatrix} \begin{pmatrix} \vdots \\ X_{-1} \\ X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{T-1} \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{T-1} \\ \epsilon_T \end{pmatrix}. \quad (4.9)$$

Equation (4.9) contains still too many parameters. In fact the number of the unknown parameter is still larger than that of observations (see the last row of the coefficient matrix). One may impose restrictions on the sequence of parameter matrices $A_i, i = 1, 2, ...$ to make them estimable. A simpler way to constrain the parameter space is to cut the causal influence at certain lags p, by assuming that $A_i = 0$ for $i > p$. This assumption implies that the causal dependence is not infinite. That is, at least from the practical point of view, an acceptable simplification. For $p = 2$ the causal model is written as follows.

$$\begin{pmatrix} A_0 & 0 & \dots & & \dots & 0 \\ A_1 & A_0 & 0 & & \dots & 0 \\ A_2 & A_1 & A_0 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & A_2 & A_1 & A_0 & 0 \\ 0 & \dots & 0 & A_2 & A_1 & A_0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{T-1} \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{T-1} \\ \epsilon_T \end{pmatrix}. \quad (4.10)$$

The first two rows represent the initial condition for the time series. The other rows represent the invariant causal relations over time.

Based on the discussion above, we will define the time series causal models(TSCM) as follows.

Definition 4.1 (TSCM) *A linear recursive model of time series is called a time-series-causal-model if it satisfies the following three constraints:*

- *temporal causal constraint,*

- *causal time-invariant constraint, and*
- *finite causal influence constraint.*

Besides the initial condition the matrix equation (4.10) can be written as follows:

$$A_0X_t + A_1X_{t-1} + A_2X_{t-2} = \epsilon_t, \qquad t = p+1, ..., T \tag{4.11}$$

where $E(\epsilon_t\epsilon'_{t-s}) = 0$, $E(\epsilon_t\epsilon'_t) = D$ and D is a diagonal matrix. The causal relations among the time series variables are expressed by the coefficient matrices $A_0, A_1, A_2, ..., A_p$. A_0 is itself a lower triangular matrix. It describes the contemporaneous causal relations among the elements of the n-vector X_t. A_i describes the causal dependence between the elements of X_t and elements of X_{t-i}. Zero elements in the coefficient matrices A_i correspond to missing edges in the DAG and hence implies no direct causal influence.

4.2 Granger Causality vs. the Probabilistic Causality

Although TSCMs as defined above are based on the fundamental assumption of the representation of causal relation in DAGs that are equivalent to the recursive simultaneous equations models in linear cases, there is an intimate formal relation between TSCMs and VAR models of time series.

Proposition 4.1 (TSCM and VAR)

Under the assumption of homoscedasticity, a TSCM has a VAR representation. A VAR corresponds to a TSCM.

Proof: A VAR model is denoted as follows:

$$X_t = \sum_{i=1}^{p} \Pi_i X_{t-i} + U_t. \qquad \text{for } t = p+1, p+2, ..., T, \tag{4.12}$$

and $E(U_tU'_t) = \Sigma$. Without loss of generality we take $p = 2$.

Premultiply the inverse of A_0 to both sides of the equation (4.11) we get:

$$X_t = -A_0^{-1}A_1X_{t-1} - A_0^{-1}A_2X_{t-2} + A_0^{-1}\epsilon_t, \qquad t = p+1, ..., T. \tag{4.13}$$

We have $E(A_0^{-1}\epsilon_t\epsilon'_tA_0^{-1'}) = A_0^{-1}D'A_0^{-1'}$. Under the assumption of homoscedasticity we have:$\Sigma := A_0^{-1}DA_0^{-1'}$. It follows that the Equation (4.13) is a VAR(p) model.

On the other hand, for any covariance matrix Σ of a VAR model like (4.12) there exists at least one decomposition, for instance the triangular decomposition, such that the following holds:

$$A_0^* \Sigma A_0^{*\prime} = S, \tag{4.14}$$

where A_0^* is a lower triangular matrix and S is a diagonal matrix. Premultiplying (4.12) by the inverse of ${A_0}^*$, we obtain:

$$A_0^* X_t - \sum_{i=1}^{p} A_0^* \Pi_i X_{t-i} = A_0^* U_t. \tag{4.15}$$

Since $A_0^* U_t$ has a diagonal covariance matrix, its components are independent. Obviously, together with the initial condition, (4.15) is formally a TSCM.

□

In the context of time series analysis one often used concept is the Granger causality. Given the correspondence between TSCMs and VAR models, it is of interest to describe the relation between the Granger causality and the causal dependence implied in a TSCM. Generally, the Granger causality and the causal dependence are two different concepts: While the Granger causality describes the relation between an element of an n-vector X_t, say $X_{i,t}$ and whole sequence of other elements of time series $X_{j,t-s}$ for all $s > 0$, the probabilistic causal dependence describes the relation between two single elements $X_{i,t}$ and $X_{j,s}$. However, in the VAR framework, the Granger causality can be formulated as zero restrictions on the parameters of a VAR model and the probabilistic causal dependence can also be presented by a zero restriction on a TSCM. Using the correspondence between TSCM and VAR we get the following relations between the Granger causality and the probabilistic causal dependence.

Proposition 4.2 (Granger Causality and TSCM)

Let $pX_{i,t}$ denote all the elements of X_t that are predecessors of $X_{i,t}$ in the TSCM. If the elements $X_{k,t-s}$ for $s = 0, 1, 2, ...p$ does not have temporal causal influence on $pX_{i,t}$ and $X_{i,t}$, then $X_{k,t}$ does not Granger cause $X_{i,t}$.

Proof: Given the correspondence between VAR (4.12) and TSCM (4.11), we have the relation

$$\Pi_s(i,k) = \sum_{j=1}^{n} A_0^{(-1)}(i,j) A_s(j,k), \tag{4.16}$$

where $\Pi_s(i,k)$ is the (i,k) element of the VAR coefficient matrix Π_s, $A_0^{(-1)}(i,j)$ and $A_s(j,k)$ are the (i,j) and (j,k) element of the TSCM coefficient matrices

A_0^{-1} and A_s, respectively. In the VAR framework the non-Granger causality of $X_{k,t}$ for $X_{i,t}$ means $\Pi_s(i,k) = 0$ for $s = 1, 2, ...p$. Because A_0 is a lower triangular matrix the inverse of A_0 is also a lower triangular matrix. We have

$$\Pi_s(i,k) = \sum_{j=1}^{n} A_0^{(-1)}(i,j)A_s(j,k) = \sum_{j=1}^{j} A_0^{(-1)}(i,j)A_s(j,k) = 0. \qquad (4.17)$$

The last equation follows from the assumption that $X_{k,t-s}$ does not have any causal influence on $pX_{i,t}$ and $X_{i,t}$. □

Remark: The following example shows that no probabilistic causal dependence of $X_{i,t}$ on any X_{jt-s} for $s \geq 0$ is not enough to ensure that X_{jt} does not Granger cause X_{it}.

$$A_1 = \begin{pmatrix} -0.4 & 1.7 & -2.2 & -0.0 \\ 0.2 & 0.4 & 1.0 & 0.8 \\ 1.0 & 0.0 & -0.8 & 1.6 \\ -0.1 & 0.1 & -0.8 & 2.1 \end{pmatrix} A_0 = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 1.0 & 0.0 \\ 0.2 & 1.0 & 0.6 & 1.0 \end{pmatrix} \qquad (4.18)$$

$$\Pi = A_0^{-1}A_1 = \begin{pmatrix} -0.4 & 1.7 & -2.2 & -0.0 \\ 0.7 & -1.4 & 3.3 & 0.8 \\ 1.4 & -1.7 & 1.3 & 1.6 \\ -1.8 & 2.2 & -4.5 & 0.1 \end{pmatrix} \qquad (4.19)$$

$A_0(3,2) = 0$ and $A_1(3,2) = 0$ imply that X_{2t} has neither contemporaneous nor temporal causal influence on X_{3t}. But $\Pi(3,2) \neq 0$ implies that X_{2t} Granger causes X_{3t}. A TSCM measures the direct effect of a variable on the other. To make an optimal prediction of a variable, say X_{it}, in a TSCM we have to known the values of all its parents. Conditional on knowing the values of these parents the values of other variables are irrelevant for the prediction. This is expressed by the zero coefficients before those variable that are not in the set of the parents of X_{it}. However, if we do not know the values of the parents of X_{it}, the values of non-parent variables may be useful to predict the values of these parents. In this case knowing the value the the non-parents variables may improve the prediction. This is why the variables X_{kt-s} with $s = 0, 1, 2, 3, ..P$ are not probabilistic causes of X_{it} but X_{kt} may Granger cause X_{it}, i.e. the X_{kt-s} with $s = 1, 2, 3, ..P$ may be useful for prediction of X_{it}.

4.3 Learning TSCMs

Similar to the cases of causal models for cross sectional data, the most important issue of the statistical treatment of TSCMs is whether we can recover

the underlying causal TSCM if the data are generated by the TSCM. We could directly apply those algorithms developed for the independent data, if we had repeated observations on the same time series. But the typical situation in economics is that we have only one observation at each point in time.

Our strategy is a two step procedure[16]: We infer the contemporaneous causal structure first. In the second step we infer the temporal causal structure. Concretely we estimate an unconstrained VAR model for the data to obtain consistent estimates of the residuals. These estimated residuals can be used as input data to learn the contemporaneous structure of A_0. The learning of the contemporaneous causal structure A_0 can be done by using the methods described in the previous section. After we get an estimate for the contemporaneous causal structure A_0, i.e. the zero restrictions on the A_0 matrix as well as identifying the order of the variables, we have a recursive SEM. We can use the BIC criterion to select models over all subsets of the lagged variables and hence determine ${A_i}^*$.

$$A_0^0 X_t + \sum_{i=1}^{p} A_i^* X_{t-1} = \epsilon_t, \tag{4.20}$$

where A_0^0 is the contemporaneous causal structural matrix with zero restrictions identified in the first step and A_i^* is the uncovered temporal causal structure coefficient using BIC.

Proposition 4.2 *[Two step procedure for TSCMs]*

- *If the contemporaneous causal structure of the data generating TSCM is observationally distinguishable, the two step procedure will identify the true causal structure of the TSCM consistently.*
- *If a TSCM is observationally distinguishable but the contemporaneous causal structure is observationally indistinguishable, the two step procedure with a consistent model selection criterion will "uniquely" identify the data generating causal model consistently.*
- *If a TSCM is observationally indistinguishable, then the two step procedure with a consistent model selection criterion will uniquely identify the causal order of the simultaneous causal blocks.*

Proof: By applying a consistent model selection criterion, we can consistently identify the true lag length of the VAR model. Because the estimate of

[16]Other approaches such as hidden Markov models or dynamic Bayesian networks can also be applied. See Kevin Murphy(1998) for details.

the covariance matrix is consistent and the true structure is observationally distinguishable, a consistent learning procedure, such as a model selection algorithm based on the BIC criterion, will identify the contemporaneous causal structure consistently[17].

We have $\underset{T\to\infty}{\text{plim}}\hat{A}_0 = A_0$. It follows that the data generating causal model is asymptotically nested in the recursive SEM (4.20). The uncovering of temporal causal structure becomes a problem of model selection in a classic regression model. As BIC criterion is consistent, we will consistently identify the temporal structure by using BIC.

If contemporaneous causal structure is not observationally distinguishable, then in the first step we can only consistently identify a class of contemporaneous causal structures that are observationally equivalent to the true contemporaneous causal structure. Each member of this identified class implies a recursive SEM. Because the data-generating causal model is observationally distinguishable, searching over all members of the observational equivalent class, the one chosen by BIC criterion: A_i^* for $i = 1, 2, ..., p$ will converge to the data-generating temporal causal structure A_i for $i = 1, 2, ...p$ asymptotically.

The third case is just a restatement of the Proposition 2.9 □

4.4 Simulation Studies

In this subsection we document some simulation results. The reasons for a simulation study are the following. (1) The results in the last section are asymptotically valid. For empirical applications, the small sample properties of the procedure are more relevant. Because simulation is a convenient way to study the small sample properties in specific settings, we run simulations to assess the performance of the two step procedure. (2) Although there are some simulation results about the performance of Bayesian network models, our input for learning the contemporaneous causal structure is not independently generated random numbers, but the estimated residuals of a unconstrained VAR. Demiralp and Hoover (2004) document some simulation results of learning the causal structure from VAR residuals by the PC-algorithm. They found that the PC algorithm can recover the true structure only moderately well. We investigate here the effectiveness of the two

[17]It is well known that a consistent model selection criterion can identify the true model if the true model is within the set of the candidate models. The practical difficulty is that we can surely get the true model only at polynomial time. Therefore for large systems only heuristic procedures are applied such that we get in these cases only a local optimum but not always the global optimum.

step procedure with a local greedy search algorithm with random restarts based on BIC criterion.

Three kinds of models are considered: the first one is an observationally distinguishable TSCM with an observationally distinguishable contemporaneous causal structure. For this kind of model we can learn the contemporaneous causal structure first, and then the temporal causal structure. The second model is an observationally distinguishable TSCM with an observationally indistinguishable contemporaneous causal structure. For this kind of model we obtain, in the first step, a class of observationally equivalent contemporaneous causal models. For each member in the observationally equivalent class, we then use BIC criterion to search over all subsects of the lagged variables and determine the contemporaneous and temporal causal structures simultaneously. The third model is an observationally indistinguishable TSCM: we can only identify the observationally equivalent class.

For the cases of observationally distinguishable data-generating TSCMs we record the frequency of the correctly recovered true causal structure. For the cases of observationally indistinguishable data-generating TSCMs we record the frequency of the correctly recovered observationally equivalent models of the data generating causal model.

In order to evaluate the effect of the range of signal-to-noise, our parameters of the data-generating TSCM are chosen in a way, such that the expected t-statistics for these parameters, in the maximum likelihood estimates of the corresponding unconstraint SVAR (4.20), are roughly the same for A_0 and A_1 respectively. Therefore, the number of observations can be used to adjust the range of signal-to-noise. We classify the signal-to-noise strength as follows: $E(|t|) < 2$ as L(low), $2 < E(|t|) < 6$ as M (middle), $6 < E(|t|)$ as H (High).

To recover the contemporaneous causal structure we apply a greedy search algorithm with random restart to the estimated residuals of the unconstrained VAR. The greedy search algorithm looks for the best improvement of a network locally by adding an edge, removing an edge or reversing the direction of an edge. The network score is based on the BIC criterion. The algorithm stops at a local optimum. To recover the temporal causal structure for an identified contemporaneous causal structure, we use BIC criterion to select the temporal causal structure over all subsets of the lagged variables.

4.4.1 Model 1: Observationally distinguishable TSCMs with an observationally distinguishable contemporaneous causal structure

The data generating TSCM is as follows:

$$A_0 X_t = A_1 X_{t-1} + \epsilon_t, \tag{4.21}$$

with

$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix} \qquad A_1 = \begin{pmatrix} 0.7 & 0.7 & 0.7 \\ 0.5 & 0.7 & 0 \\ 0 & -0.7 & 0.7 \end{pmatrix} \qquad E(\epsilon_t \epsilon_t') = \Omega = I.$$

Obviously, this TSCM is observationally distinguishable because the zero elements of two adjacent rows in (A_1, A_0) are not in the same columns.

T	$A_0\|FS$	$A_1\|A_0, FS$	$A_0\|GS$	$A_1\|A_0, GS$	Signal
20	464	123	53	13	ML
40	491	263	401	229	HL
60	486	379	423	302	HL
80	498	414	431	355	HL
100	499	445	428	383	HL
120	499	464	431	385	HL
140	500	465	438	404	HM
160	500	482	453	437	HM

Table 1: Frequency of the correctly recovered contemporaneous and temporal causal structures A_0 and A_1 in Model 1 with 500 replications.

Table 1 records the simulation results for model 1 with 500 runs. The first column with the header T reports the number of the observations used in each simulation. The second column with the header $A_0|FS$ reports the frequency of the correctly recovered contemporaneous causal structure using BIC criterion by searching over all possible models. Here we see that if the signal level for the contemporaneous causal structure is M, that is denoted by the first letter in the last column of the table, the BIC criterion can recover the true contemporaneous causal structure only moderately well. The third column with the header $A_1|A_0, FS$ reports the frequency of the uniquely and correctly recovered temporal causal structure by using the BIC criterion for each equation in the model. The difference between the second and the third column is the number of the frequency of the cases in which the contemporaneous causal structure can be correctly identified, but the temporal causal structure cannot. If the signal level of the temporal causal structure is L, that is denoted by the second letter in the last column, we cannot get a satisfactory result. The signal of the temporal causal structure of level M or higher is enough to ensure rather good results. The fourth column with the

header $A_0|GS$ reports the frequency of the correctly identified contemporaneous causal models using greedy search. $A_1|A_0, GS$ reports the frequency of the correctly identified temporal and contemporaneous causal models using greedy search. The difference between the fourth and the fifth column is the number of the frequency of the cases in which the contemporaneous causal structure can be correctly identified, but the temporal causal structure cannot. Obviously, this algorithm can recover the true structure only moderately well, when the signal of the contemporaneous causal structure is M; the performance improves when the signal becomes H. Again, when the temporal signal is low, the temporal causal structure cannot be satisfactorily identified. The last column with the header *Signal* reports the signal-noise range of the data generating causal model. The first letter reports the signal level of the contemporaneous causal structure and the second letter reports the signal level of the temporal causal structure. Because the parameters of the data generating causal model remain unchanged in the simulation runs the increase of the number in observations leads to the increase of the strength in the signal level.

4.4.2 Model 2: Observationally distinguishable TSCMs with observationally indistinguishable instantaneous causal structure

The data generating TSCM is as follows[18]: The data generating TSCM is as follows:

$$A_0 X_t = A_1 X_{t-1} + \epsilon_t, \tag{4.22}$$

with

$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad A_1 = \begin{pmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{pmatrix} \qquad E(\epsilon_t \epsilon_t') = \Omega = I$$

In this model, the zero element of two adjacent rows are not in the same columns of (A_1, A_0). It is observationally distinguishable. But A_0 matrix contains zero elements in same columns for some adjacent rows. Therefore we do not have an observationally distinguishable contemporaneous causal structure.

[18] In the DGP here the elements on the principle diagonal of A_0 matrix are not always normalized to one. They can be normalized to one then the covariance matrix will not have unit variance. As we are only interested in identifying the causal structure, this does not make any difference.

T	$A_0\|FS$	$A_0\|GS$	$OEA_0\|FS$	$A_1\|A_0$	$A_1, A_0\|OEA_0, FS$	$A_1\|\bar{A}_0$	Signal
20	89	13	206	348	192	104	ML
40	145	17	297	464	293	30	ML
60	114	10	356	494	352	6	HL
80	105	4	378	494	378	3	HL
100	128	1	390	498	390	2	HL
120	89	0	403	499	403	3	HL
140	93	0	426	499	426	0	HM
160	104	1	432	498	432	1	HM

Table 2: Frequency of recovering the true contemporaneous and temporal causal structure A_0 and A_1 in Model 2 with 500 replications.

The second and the third columns in Table 2 show that if the data generating contemporaneous causal structure has observationally equivalent structures, it is impossible to recover the true contemporaneous causal structure directly from the VAR residuals. But the observationally equivalent contemporaneous causal structures of A_0 can be correctly identified. Further, since the TSCM is observationally distinguishable, the temporal causal information can be used to identify the contemporaneous causal structure and the temporal causal structures. The fifth column shows the frequency of the correctly identified A_1, if A_0 is correctly given. The sixth column shows that by searching over all subsets of the lagged variables for every observationally equivalent contemporaneous causal structures identified from the VAR residuals, we can identify the contemporaneous causal structure as well as the temporal causal structure. The seventh column shows the frequency of the correctly recovered A_1 if an observationally equivalent contemporaneous causal structure is given instead of A_0 itself. Obviously, given a false contemporaneous causal structure, there is no chance to recover the temporal causal structure correctly. This simulation result supports the statement in Proposition4.2.

4.4.3 Model 3: Observationally indistinguishable TSCMs with observationally indistinguishable instantaneous causal structure

The data generating TSCM is as follows:

$$A_0 X_t = A_1 X_{t-1} + \epsilon_t, \tag{4.23}$$

with

$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad A_1 = \begin{pmatrix} 0.8 & 0.8 & 0 \\ 0.8 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{pmatrix} \qquad E(\epsilon_t \epsilon_t') = \Omega = I$$

In this model, the zero element of the first and the second rows are in the same columns of (A_1, A_0). It is not observationally distinguishable.

T	$A_0\|FS$	$A_0\|GS$	$OEA_0\|FS$	$A_1\|A_0$	$A_1\|OEA_0, FS$	$A_1\|\bar{A}_0$	Signal
20	83	7	236	329	27	29	ML
40	148	16	396	466	38	24	ML
60	109	13	436	488	36	34	HL
80	106	2	452	489	25	22	HL
100	97	0	471	496	21	22	HL
120	90	2	464	499	20	15	HL
140	70	0	453	499	23	15	HM
160	90	1	470	499	12	19	HM

Table 3: Frequency of recovering the true contemporaneous and temporal causal structure A_0 and A_1 in Model 3 with 500 replications.

The simulation result for model 3 is reported in Table 3 that is constructed in the same way as the Table 2. The numbers from the second column to fifth column confirm the result in Table 2: if the contemporaneous causal structure has observationally equivalent structures, we cannot recover the contemporaneous causal structure directly from the VAR residuals. But the observationally equivalent structures can be correctly recovered, which is shown in the fourth column under the header $OEA_0|FS$. The fifth column under the header $A_1|A_0$ shows that even if A_0 is given correctly, A_1 can be correctly recovered. The numbers in the sixth column under the header $A_1|OEA_0, FS$ show the frequency of correctly recovered A_1 by searching over all subsets of the lagged variables for all the observationally equivalent structures of A_0. Obviously, in this case we cannot correctly recover A_1 since the temporal causal structure also has observationally equivalent structures.

The simulation result shows that the signal-noise range measured by the expected t-statistics in the unconstrained VAR is crucial for the performance of the learning procedure. We summarize the simulation results as follows

- When the signal of the contemporaneous causal structure is M the learning procedure can only identify the true contemporaneous structure A_0 (up to observational equivalence) moderately well. Consequently the frequency of detecting the true total structure is also only moderately often or worse. If the signal of the contemporaneous causal structure is H, then the true contemporaneous structure can be identified with high accuracy.

- The signal of the temporal causal structure on level M is enough to ensure very good performance of the learning procedure. However, if the signal of the temporal causal structure is L, the performance of the learning procedure will be negatively influenced.

- The greedy search procedure performs, generally, only moderately well. Even when the signal level for the contemporaneous causal structure is very high, the greedy search can only uncover the true contemporaneous causal structure A_0 at a relative frequency of 75%[19]. So repeated random restart is necessary to make sure that the procedure will give relatively good results.

5 An Application of the Causal Analysis to Wage-Price Dynamics

There is a view among economic professionals that higher wages lead to higher prices. The reasoning behind this view seems to be closely related to that behind the concept of Phillips curves[20] and the notion of NAIUR. Layard, Nickell and Jackman (1994) describe the reasoning as follows: "[...] when buoyant demand reduces unemployment (at least relative to recent experienced levels), inflationary pressure develops. Firms start bidding against each other for labour, and workers feel more confident in pressing wage claims. If the inflationary pressure is too great, inflation starts spiraling upwards: higher wages lead to higher price rises,[...]". Beside this intuitively appealing argument, the market-up pricing by firms provides another explanation. "Since labor costs are a large fraction of a firms total costs of production, an increase in wages and compensation should put pressure on firms to pass through these higher costs onto higher prices."

However, as argued by Hess and Schweitzer (2000) "This story is incomplete, however, for a few reasons. First, an increase in wages will not create infla-

[19]Here we confirm the results as found in Demiralp and Hoover (2004) for the PC algorithm.

[20]The original Phillips curve is expressed as empirical law of the relation between the unemployment rate and the wage inflation. See Phillips (1958)

tionary pressure if the increase in wages is brought about by increased labor productivity. Hence, controlling for labor productivity (i.e. supply effects) in the analysis between wages and prices would seem very important. Second, an increase in wages will not create inflationary pressure if the increase in wages leads to a squeeze in a firms profits due to their inability to pass along cost increases. No firm inherits the right to simply mark-up the prices of its output as a constant proportion above their costs, as competitive market pressures provide a strong influence on the pricing decisions of firms." Jonsson and Palmqvist (2004) show in a two sector general equilibrium model that wage increases do not lead to inflation.

Many economists try to clarify the controversy with the help of empirical evidence extracted by econometric methods. In the econometric literature this issue is typically translated into the question whether the wage inflation Granger-causes the price inflation. According to Hess and Schweitzer (2000) most studies have not found any strong indications that this is the case. Examples of such studies are Hogan (1998), Rissman (1995), Clark (1997) and Mehra (1993). Staiger, Stock, and Watson (1997) find that price predicts wage better than the other way around. Ghali (1999) finds strong evidence that wages Granger-cause prices based on a multivariate cointegration analysis. Aaronson (2001) finds that restaurant prices generally rise with changes in the wage bill. The empirical evidence is thus mixed.

Facing these controversial theoretical arguments and the mixed empirical evidence identified by Granger causality test so far, we are going to contribute to this issue with a new methodology of the inferred causation. "Higher wages lead to higher prices" is essentially a statement about a causal relation that implies not only the dependence but also the directionality of the dependence.

Using the methodology developed in the last section we analyze the causal dependence among the variables in a wage-price dynamic as in Chen, Chiarella, Flaschel, and Semmler (2005). There are 6 variables in this dynamic system: (dw, dp, e, u, dz, pim) are the wage inflation, the price inflation, the labor utilization rate, the capacity utilization rate, the growth of the labor productivity, and the inflationary climate, respectively. The main concern of this exercise is to demonstrate how the method of the causal analysis can be used to answer the question whether the wage inflation causes the price inflation or the price inflation cause the wage inflation.

The empirical data for the relevant variables discussed above are taken from Economic Data - FRED©[21]. The data shown below are quarterly, seasonally adjusted, annualized where necessary and are all available from 1947:1 to 2004:4. Up to the rate of unemployment they represent the business sector of the U.S. economy. We will make use in our estimations below of the range

[21] http://research.stlouisfed.org/fred2/.

from 1965:1 to 2004:4 solely, i.e., roughly speaking of the last five business cycles that characterized the evolution of the U.S economy. We thus neglect the evolution following World War II to a large degree.

Variable	Transformation	Mnemonic	Description of the untransformed series
e	log(1-UNRATE/100)	UNRATE	Unemployment Rate (%)
u	log(GDPC1/GDPPOT)	GDPC1,GDPPOT	GDPC1: Real Gross Domestic Product of Billions of Chained 2000 Dollars, GDPPOT: Real Potential Gross Domestic Product of Billions of Chained 2000 Dollars, u:Capacity Utilization: Business Sector (%)
w	log(HCOMPBS)	HCOMPBS	Business Sector: Compensation Per Hour, Index 1992=100
p	log(IPDBS)	IPDBS	Business Sector: Implicit Price Deflator, Index 1992=100
z	log(OPHPBS)	OPHPBS	Business Sector: Output Per Hour of All Persons, Index 1992=100
π_m	MA(dp)		inflationary climate measured by the moving average of price inflation in the last 12 periods

Table 4: Raw Data used for empirical investigation of the model

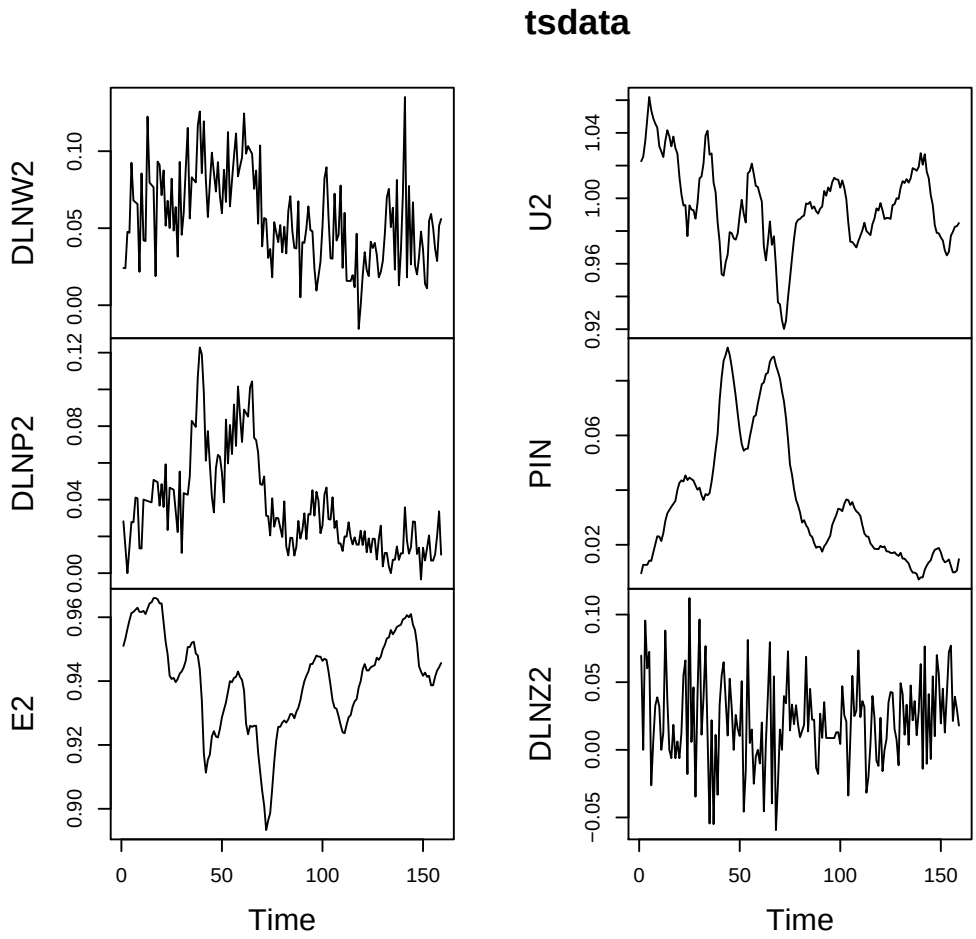

Figure 2: Data for the analysis of wage-price spiral

Before we start with our empirical investigation, we examine the stationarity of the relevant time series. The shown graphs of the series for wage and price inflation, capacity utilization rates and labor productivity growth suggest the stationarity of the time series (as expected). In addition we carry out the augmented DF unit root test for each series. The test results are reported in Table 4. The unit root tests confirms our expectation.

Variable	Sample	Critical value	Test Statistic
dw	1947:02 TO 2004:04	-3.45	-7.12
dp	1947:02 TO 2004:04	-3.45	-4.60
e	1947:02 TO 2004:04	-3.45	-4.35
u	1947:02 TO 2000:04	-3.45	-4.01
dz	1947:02 TO 2004:04	-3.45	-15.26

Table 5: Summary of DF-Test Results.

We first construct a six dimensional VAR model for (dw, dp, e, u, dz, pi_m). Using the Schwarz information criterion we select the lag length 1[22]. The Granger causality tests in this VAR(1) setting give the following results.

```
                F-statistic      p-value
dp -> dw        31.09595490    1.066199e-07
dw -> dp        6.91532076     9.407467e-03
```

We see here dw Granger causes dp and dp Granger causes dw. As discussed in the last section, these results do not give us a clear answer about whether the wage inflation leads to price inflation or the other way around.

Applying the greedy search algorithm with random restarts to the estimated residuals of the unconstrained VAR(1), we get the following DAG for the contemporaneous causal structure.

[22]The choice of one lag in a system with quarterly data seems to be very unusual. Taking into account that the inflationary climate variable π_m is a summary of the lagged information, this choice would not be so surprising. See Appendix for details about the possibility of an alternative choice of lag length.

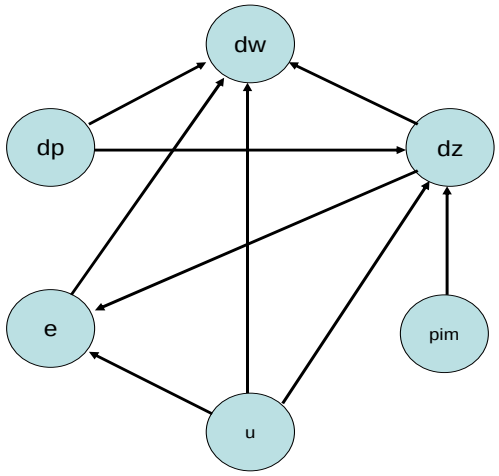

Figure 3: The contemporaneous causal graph in the wage price spiral

The corresponding contemporaneous causal structure matrix is:

$$\begin{pmatrix} 1 & -0.47 & -1.93 & 1.56 & 0 & -0.50 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -0.36 & 0 & 0.05 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.38 & 0 & -2.73 & -3.18 & 1 \end{pmatrix} \tag{5.24}$$

It is important to emphasize that all the arrows in the causal graph above are inferable. It is easy to check that the four arrows $dp \to dw$, $e \to dw$, $u \to dw$ and $dz \to dw$ constitute v-structures and therefore their directions are inferable. The three arrows $dp \to dz$, $u \to dz$ and $\pi_m \to dz$ constitute also v-structures and their directions are inferable too. Now the change of the direction of the arrow $dz \to e$ will lead to new v-structures therefore the direction of this arrow is also inferable. And the change of the direction of the arrow $u \to e$ would lead to cycle. Since all directions of the arrows are inferable, the contemporaneous causal structure identified above is observationally distinguishable and the arrows imply causal directions.

According to the causal graph, the causal order in the contemporaneous innovation is $(u, dp, \pi_m, dz, e, dw)$. This causal order corresponds to the intuition that the adjustment of the capacity utilization and the preis adjustment lead the price inflation climate and the productivity growth, and the labor utilization and the wage adjustment follow. $dp \to dw$, $e \to dw$, $u \to dw$, $dz \to dw$ implies that the wage inflation is caused contemporaneously by price inflation, the labor utilization, capacity utilization and the growth of the labor

productivity. This corresponds to often used wage Phillips curve[23].

After rearranging the contemporaneous causal structure in this order we get the recursive contemporaneous causal structure matrix:

$$A_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -2.73 & 0.38 & -3.18 & 1 & 0 & 0 \\ -0.36 & 0 & 0 & 0.05 & 1 & 0 \\ 1.56 & -0.47 & 0 & -0.50 & -1.93 & 1 \end{pmatrix} \tag{5.25}$$

In the second step we learn the temporal causal structure A_1 by applying OLS to the recursive SEM with the identified contemporaneous causal structural A_0. After neglecting the insignificant coefficient in the OLS estimation we obtain the estimated temporal causal structure:

$$A_1 = \begin{pmatrix} -1.03 & 0 & 0 & 0 & 0.28 & 0 \\ -0.19 & -0.52 & -0.40 & 0.07 & 0 & 0 \\ 0.02 & -0.12 & -0.90 & 0 & -0.08 & 0 \\ 0.64 & 0 & 0 & 0 & 0 & -0.21 \\ 0.30 & 0.02 & 0 & 0 & -0.91 & 0 \\ -2.01 & 0 & -0.54 & 0 & 1.94 & 0 \end{pmatrix} \tag{5.26}$$

According to these two matrices we can draw the causal graph for the time series data of the wage price dynamic. We observe that dp has two channels of direct contemporaneous causal influence on dw that are depicted by the red arrows. In addition it has three channels of temporal indirect causal influence on dw that are depicted by the three pink dotted arrows. dw has neither contemporaneous direct causal influence on dp nor the temporal indirect causal influence. The feedback of dw on dp goes through three periods. $dw_{t-2} \rightarrow dz_{t-1} \rightarrow dp_t$. This is represented by the green dotted line.

[23]See **?** and **?**.

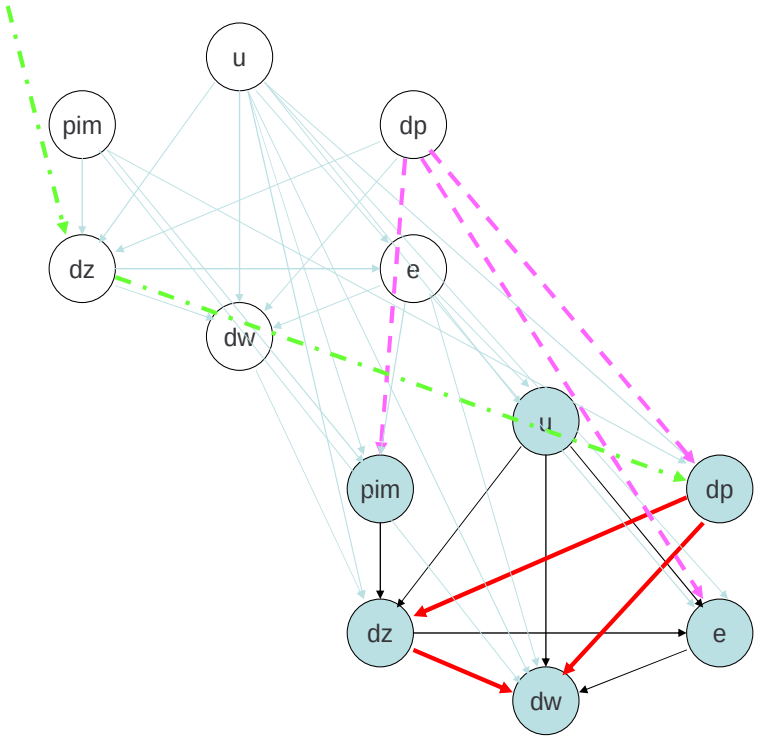

Figure 4: Causal graph of the wage-price spiral system

This derived causal structure provides a clear answer to the causal effect-relation between the price inflation and the wage inflation. The price inflation is the driving force in the wage price dynamics.

From the TSCM we obtain two structural Phillips curves, one for the price inflation one for the wage inflation as follows.

$$dp_t = 0.52dp_{t-1} + 0.40\pi_{mt-1} + 0.19u_{t-1} - 0.07dz_{t-1} - 0.21 \tag{5.27}$$

$$dw_t = 0.47dp_t + 0.54\pi_{mt-1} + 0.44u_t + 0.5dz_t + 2.0(\Delta e_t - \Delta u_t) - 0.42 \tag{5.28}$$

Unlike most formulations of Phillips curves that are derived based on theoretical arguments, these two structural Phillips curves are the results of the data-driven causal analysis. They represent the causal influence of the right-hand-side variables on the dependent variables.

The price Phillips Curve shows that the price inflation is driven by the demand pressure measured by the utilization rate of capacity u_{t-1}, and the

inflationary climate π_{mt-1} in which the economy operates i.e. the inertia of the price inflation rate. The growth of the labor productivity reduces the wage cost and hence the price inflation.

In the wage Phillips curve the wage inflation rate is driven by the demand pressure term measured by u_t, the living cost pressure measured by p_t and the inflationary climate π_{mt-1}. The growth of labor productivity acts positively on the wage inflation rate. The variable u_t is identified as a proper measure of demand pressure. Since u_t is highly correlated with the rate of labor utilization of the employed labor, this implies that the level of the rate of labor utilization of the insiders on the labor market within firms is the demand pressure that acts on wage inflation. However if the increase of labor utilization spills over from the insiders to the outsiders $\Delta e_t - \Delta u_t > 0$, large wage inflation will be expected.

6 Concluding Remarks

In this paper we develop a method to uncover the causal order in stationary multivariate time series with a vector autoregressive presentation. Complex directionality of dependence among economic variables, such as the unidirectional dependence, the simultaneous dependence, the contemporaneous dependence as well as the temporal dependence can be presented in TSCMs. A two step learning procedure is developed to uncover the potential causal relations implied in the data. This two step procedure reduces largely the dimension of the Bayesian network that is used to present the causal relations. In case of high signal-to-noise ratio in the contemporaneous causal structure, this two step procedure can effectively uncover the underlying causal structure.

The TSCMs developed in this paper can be applied to analyze the dynamic causal relations among economic variables which are of great interest for economists. The two step learning procedure for TSCMs can be used to uncover the directionality of dependence in the data, such as contemporaneous dependence as well as the temporal dependence.

We applied the TSCM to a wage-price dynamic and obtained the result that the price-inflation rate is one of the causes that drive the wage-inflation rate while the wage-inflation rate has only a very weak indirect influence on the price-inflation rate. From the TSCM of the wage-price dynamic we obtained two structural Phillips curves that represent the causal influence in the determination of the price-inflation rate and the wage-inflation rate. As structural equations in economics are genuinely interpreted as causal relation, TSCMs provide a way to derive structural equations in which the causal interpretation of the relations is justified.

As the application of the method of inferred causation for the identification of causal relations among economic variables is still fairly new, many issues such as the robustness of the resulting causal graphs with respect to the choice of different sample periods, implications of relaxing the triangularity assumption on A_0, the influence of the applied statistical criteria in the learning procedure, the efficiency of the algorithm, or the technique for obtaining a structure that is globally optimal, deserve further investigation.

7 Appendix

- Observational Equivalence $\Leftrightarrow$ same Ω up to permutation of the variables.
- Let $\mathcal{O}$ be one order on $X = \{x_1, \cdots, x_n\}$. Every order corresponds to one unique triangular decomposition of $A_{\mathcal{O}}$ (lower triangular). $A_{\mathcal{O}}X$ has orthogonal innovations and it is called a *causal model*.
- By changing a given order $\mathcal{O}$ on X to another order $\mathcal{O}'$, usually, the number of null restrictions in $A_{\mathcal{O}'}$ will reduce. Let $\mathcal{O}_0$ represent the order with most null restrictions (with respect to given Ω).

Lemma 7.1 *Let x, y be random variables with $\mathbf{E}[x] = 0$, $\mathbf{E}[y] = 0$. Let $z, z_1, \cdots, z_m$ be random variables and $Z = \{z_1, \cdots, z_m\}$. Let $x|z$ be conditional variable of x on z. Let $\mathcal{P}(x|z)$ be the linear projection of x on z, then*

$$x|z = x - \mathcal{P}(x|z) \ .$$

The conditional correlation has the following equality

$$\begin{aligned} & \mathbf{Cov}[x|\{Z \cup z\}, y|\{Z \cup z\}] \qquad (7.29) \\ = \ & \mathbf{Cov}[x|Z, y|Z] - \mathbf{Cov}[x|Z, z|Z]\mathbf{Var}[z|Z]^{-1}\mathbf{Cov}[z|Z, y|Z] \ . \end{aligned}$$

Proof of Proposition 2.5
To prove Proposition 2.5 we show that if an interchange of two variables keeps the number of the null restrictions, then these two variables must be sequential neighbors and the corresponding rows in A must have zeros in same columns.

The coefficients a_{ij} in A can be interpreted as the partial regression coefficient

$$a_{ij} = r_{x_i, x_j|X_{i-1,\hat{j}}} = \mathbf{Cov}[x_i, x_j|X_{i-1,\hat{j}}]/\mathbf{Var}[x_j|X_{i-1,\hat{j}}] \ ,$$

where $X_{i-1} = \{x_1, \cdots, x_{i-1}\}$, $\hat{j}$ means exclusion of x_j, $j < i$.

Consider at first two orders $\mathcal{O} = \{1, 2, \cdots, k, k+1, k+2, k+3, \cdots, n\}$ and $\mathcal{O}' = \{1, 2, \cdots, k, k+2, k+1, k+3, \cdots, n\}$ where the positions of x_{k+1} and x_{k+2} are exchanged. Let a_{ij} and a^*_{ij} represent the triangular coefficients with respect to the order $\mathcal{O}$ and $\mathcal{O}'$.

From the interpretation of the triangular coefficients we have

$$\begin{aligned}
a_{k+1,j} = r_{x_{k+1},x_j|X_{k,\hat{j}}} &= \mathbf{Cov}[x_{k+1}, x_j|X_{k,\hat{j}}]/\mathbf{Var}[x_j|X_{k,\hat{j}}] && (7.30)\\
a_{k+2,j} = r_{x_{k+2},x_j|X_{k+1,\hat{j}}} &= \mathbf{Cov}[x_{k+2}, x_j|X_{k+1,\hat{j}}]/\mathbf{Var}[x_j|X_{k+1,\hat{j}}] && (7.31)\\
a^*_{k+1,j} = r_{x_{k+2},x_j|X_{k,\hat{j}}} &= \mathbf{Cov}[x_{k+2}, x_j|X_{k,\hat{j}}]/\mathbf{Var}[x_j|X_{k,\hat{j}}] && (7.32)\\
a^*_{k+2,j} = r_{x_{k+1},x_j|(X_{k,\hat{j}}\cup,x_{k+2})} &= \mathbf{Cov}[x_{k+1}, x_j|(X_{k,\hat{j}} \cup x_{k+2})]/\mathbf{Var}[x_j|(X_{k,\hat{j}} \cup x_{k+2})] && (7.33)
\end{aligned}$$

Applying Lemma 7.1 on the equalities above we can obtain the following equalities, for $j \leq k$,

$$\begin{aligned}
a^*_{k+1,j} &= a_{k+2,j}\ \phi_{11} + a_{k+1,j}\ \phi_{12} && (7.34)\\
a^*_{k+2,j} &= a_{k+1,j}\ \phi_{21} - a^*_{k+1,j}\ \phi_{22}\ , && (7.35)
\end{aligned}$$

where

$$\phi_{11} = \frac{\mathbf{Var}[x_j|X_{k+1,\hat{j}}]}{\mathbf{Var}[x_j|X_{k,\hat{j}}]} > 0\ , \qquad \phi_{12} = \frac{\mathbf{Cov}[x_{k+2}, x_{k+1}|X_{k,\hat{j}}]}{\mathbf{Var}[x_{k+1}|X_{k,\hat{j}}]}$$

$$\phi_{21} = \frac{\mathbf{Var}[x_j|X_{k,\hat{j}}]}{\mathbf{Var}[x_j|X_{k,\hat{j}} \cup x_{k+2}]} > 0\ , \qquad \phi_{22} = \frac{\mathbf{Cov}[x_{k+1}, x_{k+2}|X_{k,\hat{j}}]}{\mathbf{Var}[x_{k+2}|X_{k,\hat{j}}]}\ .$$

Using these two equalities above we can have the following results easily

$$\begin{aligned}
(a_{k+1,j} = 0, a_{k+2,j} = 0) &\Rightarrow (a^*_{k+1,j} = 0, a^*_{k+2,j} = 0) && (7.36)\\
(a_{k+1,j} \neq 0, a_{k+2,j} \neq 0) &\Rightarrow (a^*_{k+1,j} \neq 0, a^*_{k+2,j} \neq 0) && (7.37)\\
(a_{k+1,j} = 0, a_{k+2,j} \neq 0) &\Rightarrow (a^*_{k+1,j} \neq 0, a^*_{k+2,j} \neq 0) && (7.38)\\
(a_{k+1,j} \neq 0, a_{k+2,j} = 0) &\Rightarrow (a^*_{k+1,j} \neq 0, a^*_{k+2,j} \neq 0) && (7.39)
\end{aligned}$$

and with these results we conclude that the positions of null constraints have to be the same in the two interchanged neighboring rows.

We have to remark that we exclude the cases of in which the parameter may assume a particular value zero (faithfulness assumption). For example, that $a_{k+2,j}\ \phi_{11} + a_{k+1,j}\ \phi_{12}$ occasionally equal to zero while $a_{k+2,j} \neq 0$ and $a_{k+1,j} \neq 0$.

We now prove the equivalence between (V-structure + Skeleton) condition and our interchange condition. First we prove our interchange condition keeps the skeleton and all V-structure of the graph. $k+1$-th and $k+2$-th rows have the same positions of zeros means x_{k+1} and x_{k+2} always have the same parents, say $x_i, i \leq k$. For changing the order of x_{k+1}, x_{k+2}, the

positions of zeros remain unchanged, i.e. the set of the common parents in the graphic remains unchanged. If $x_{k+1} \to x_{k+2}$ exists ($a_{k+2,k+1} \neq 0$),

$$\begin{array}{ccc} x_i & \longrightarrow & x_{k+1} \\ & \searrow & \downarrow \\ & & x_{k+2} \end{array},$$

the interchange turn the arrow between x_{k+1} and x_{k+2}. As shown, the interchange cannot take place in a V-structure since x_{k+1} and x_{k+2} have always the same parents.

If there is no arrow between x_{k+1} and x_{k+2} ($a_{k+2,k+1} = 0$), the interchange does not have any effect on the graphics. So we proved the interchange rule keep the skeleton and V-structure.

We prove now the keeping of skeleton and V-structures can only be done by our interchange condition.

To keep a given skeleton structure while exchanging two rows which are not neighbors, for example, $k+1$-th and $k+3$-th rows

	$\ddots$	x_{k+1}	x_{k+2}	x_{k+3}	$\cdots$
x_{k+1}	$\cdots$	1	0	0	$\cdots$
x_{k+2}	$\cdots$	$a_{k+2,k+1}$	1	0	$\cdots$
x_{k+3}	$\cdots$	$a_{k+3,k+1}$	$a_{k+3,k+2}$	1	$\cdots$
$\vdots$					$\ddots$

$$\Downarrow$$

	$\ddots$	x_{k+3}	x_{k+2}	x_{k+1}	$\cdots$
x_{k+3}	$\cdots$	1	0	0	$\cdots$
x_{k+2}	$\cdots$	$a^*_{k+2,k+1}$	1	0	$\cdots$
x_{k+1}	$\cdots$	$a^*_{k+3,k+1}$	$a^*_{k+3,k+2}$	1	$\cdots$
$\vdots$					$\ddots$

Table 4:

it is necessary to have $a^*_{k+2,k+1} = 0$ because the relation $x_{k+3} \to x_{k+2}$ does not exist in the old graphic. And if $a^*_{k+2,k+1} = 0$, the interchange of the $k+1$-th row (corresponding x_{k+3}) and the $k+2$-th row in the A-matrix represents the identical graphic. Therefore, we can interchange x_{k+3} and x_{k+2} in the A-matrix

	$\ddots$	x_{k+2}	x_{k+3}	x_{k+1}	$\cdots$
x_{k+2}	$\cdots$	1	0	0	$\cdots$
x_{k+3}	$\cdots$	0	1	0	$\cdots$
x_{k+1}	$\cdots$	$a^{**}_{k+3,k+1}$	$a^{**}_{k+3,k+2}$	1	$\cdots$
$\vdots$					$\ddots$

.

Table 5:

Similarly we have $a^{**}_{k+3,k+2}$ in Table 5 equal to zero because the causation relation $x_{k+3} \longrightarrow x_{k+1}$ does not exist in the old graphic. Therefore we can exchange the $k+2$-th row (corresponding x_{k+3}) and $k+3$-th row (corresponding x_{k+1} and obtain

	$\ddots$	x_{k+2}	x_{k+1}	x_{k+3}	$\cdots$
x_{k+2}	$\cdots$	1	0	0	$\cdots$
x_{k+1}	$\cdots$	0	1	0	$\cdots$
x_{k+3}	$\cdots$	$a^{***}_{k+3,k+1}$	0	1	$\cdots$
$\vdots$					$\ddots$

.

Table 6:

All together means that the interchange between x_{k+1} and x_{k+3} under the maintain of the skeleton structure can be also done by interchanging at first x_{k+1} and x_{k+2} and then interchanging x_{k+1} (now in $k+2$-th row) and x_{k+3} which are in neighboring rows.

Now we prove the maintain of V-structure is followed only by the position constraints in the neighboring rows. When we turn the arrow between x_{k+1} and x_{k+2} and assume there exists $i \leq k$ such that $a_{j,i} = 0$ and $a_{j+1,i} \neq 0$. Then, we have a V-structure on the set $\{x_i, x_{k+1}, x_{k+2}\}$ as shown

$$\begin{array}{ccc} x_i & & x_{k+1} \\ & \searrow & \downarrow \\ & & x_{k+2} \end{array} \quad \Rightarrow \quad \begin{array}{ccc} x_i & & x_{k+1} \\ & \searrow & \uparrow \\ & & x_{k+2} \end{array} ,$$

which is changed after interchange. It contradicts to the maintain of the V-structure. So the positions of zeros in these two rows have to be the same.

References

Aaronson, D. (2001). Price pass-through and the minimum wage. *The Review of Economic Studies*, 83:158–169.

BACH, F. R. AND JORDAN, M. I. (2004). Learning graphical models for stationary time series. *IEEE Tansmitions on signal processing*, 52:2189–2199. *Further information*

CARTWRIGHT, N. (2001). What is wrong with bayes nets? *MONIST*, 84:242–264. *Further information*

CHEN, P., CHIARELLA, C., FLASCHEL, P., AND SEMMLER, W. (2005). Keynesian macrodynamics and the phillips curve. an estimated baseline macro-model for the U.S. economy. Quantitative and Empirical Analysis of Nonlinear Dynamic Macromodels. *Further information in IDEAS/RePEc*

CLARK, T. (1997). Do producer prices help to predict consumer prices? *Federal Reserve Bank of Kansas City Research Paper*, No.97-09. *Further information in IDEAS/RePEc*

DAHLHAUS, R. (2000). Graphical interpretation for multivariate time series. *Metrika*, 51:157–172.

DEMIRALP, S. AND HOOVER, K. (2004). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65:745–767. *Further information*

DHRYMES, P. J. (1993). *Topics in Advanced Econometrics*. Springer-Verlag, 1st edition.

EICHLER, M. (2003). Granger causality and path diagrams for multivariate time series. *Discussion papers, Department of Statistics, the University of Chicago. Further information in IDEAS/RePEc*

FREEDMAN, D. AND HUMPHREYS, P. (1998). Are there algorithms that discover causal structure? *http://www.stanford.edu/class/ed260/freedman514.pdf*

GHALI (1999). Wage growth and the inflation process: A multivariate cointegration analysis. *Journal of Money, Credit and Banking*, 31:417–431.

GLYMOUR, S. AND SPIRTES, G. (1988). Latent variables, causal model and overidentifying constraints. *Journal of Econometrics*, 39:175–198. *Further information in IDEAS/RePEc*

HECKERMAN (1995). A tutorial on learning with bayesian networks. Microsoft Research, MSR-TR-95-06.
ftp://ftp.research.microsoft.com/pub/tr/tr-95-06.pdf

HECKERMAN, D., GEIGER, D., AND CHICKERING, D. (1995). Learning bayesian network: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

HESS, G. AND SCHWEITZER, M. (2000). Does wage inflation cause price inflation? *Federal Reserve Bank of Cleverland Policy Discussion Paper No.10*.
Further information in IDEAS/RePEc

HOGAN, V. (1998). Explaining the recent behavior of inflation and unemployment in the united states. *IMF working paper No. 98/145*.
Further information in IDEAS/RePEc

HOOVER, K. (2005). Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory*, 21:69–77.
Further information in IDEAS/RePEc

JONSSON, M. AND PALMQVIST, S. (2004). Do higher wages cause inflation? SVERIGES RISKBANK , Working Paper Series 159.
Further information in IDEAS/RePEc

MEHRA, Y. (1993). Unit labor costs and the price level. *Federal Reserve Bank of Richmond Economic Review* , 79:25–53.
Further information in IDEAS/RePEc

PEARL, J. (2000). Causality. Cambridge University Press, 1st edition.
Further information

PEARL, J. AND VERMA, T. (1991). A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall (Eds.), Principles of Knowledge Representation and Reasoning: Procedings of the 2nd International Conference, San Mateo, CA: Morgan Kaufmann.:441-52.

PHILLIPS, A. (1958). The relation between unemployment and the rate of change of money wage rates in the united kingdom 1861–1957. *Economica*, 25:283–299.

RISSMAN, E. (1995). Sectoral wage growth and the inflation. *Federal Reserve Bank of Chicago*, Economic Perspevtives, July/August:16–28. *Further information in IDEAS/RePEc*

SPIRTES, P., GLYMOUR, C., AND SCHEINES, R. (2001). *Causation, Prediction and Search*. Springer-Verlag, New York / Berlin / London / Heidelberg / Paris, 2nd edition. *Further information*

STAIGER, D., STOCK, J. H., AND WATSON, M. W. (1997). The nairu, unemployment and monetary policy. *Jounal of Economic Perspectives*, 11:33–49. *Further information in IDEAS/RePEc*

SWANSON, N. And Granger, J. (1997). Impulse response functions based on causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92:357–367. *Further information in IDEAS/RePEc*