A time series causal model

Chen, Pu

Melbourne University

September 2010

A Time Series Causal Model

Pu Chen[*]

September 8, 2010

Abstract

Cause-effect relations are central in economic analysis. Uncovering empirical cause-effect relations is one of the main research activities of empirical economics. In this paper we develop a time series casual model to explore casual relations among economic time series. The time series causal model is grounded on the theory of inferred causation that is a probabilistic and graph-theoretic approach to causality featured with automated learning algorithms. Applying our model we are able to infer cause-effect relations that are implied by the observed time series data. The empirically inferred causal relations can then be used to test economic theoretical hypotheses, to provide evidence for formulation of theoretical hypotheses, and to carry out policy analysis. Time series causal models are closely related to the popular vector autoregressive (VAR) models in time series analysis. They can be viewed as restricted structural VAR models identified by the inferred causal relations.

JEL CLASSIFICATION SYSTEM FOR JOURNAL ARTICLES:
C1,E31

KEYWORDS: Inferred Causation, Automated Learning, VAR, Granger Causality, Wage-Price Spiral

[*]Melbourne University, E-Mail: puc@unimelb.edu.au. This research was supported by the Faculty Research Grant of Faculty of Economics and Business of Melbourne University.

Contents

1 Introduction

In the middle of the last century, Wold (1954) proposed using a recursive model structure to analyze causal relations among economic time series. His causal inquiry in economic time series encountered problems of the existence of observational equivalence and a then not-yet established theoretic justification of the use of recursive equations to represent causal structure. The research activities on this causal inquiry gave ways to the focus on identification problems in simultaneous equation systems, which could be regarded as an alternative form of articulating causal relations among economic time series variables (See Hoover (2008) for more details.).

In this paper we carry on the inquiry started in Wold (1954) and ground the recursive model structure for time series on theory of inferred causation. The theory of inferred causation is a graph-theoretic approach to causality that was first developed in the science disciplines of computer science and philosophy. A comprehensive account of this causal approach is given in Spirtes, Glymour, and Scheines (2000) and Pearl (2000). Despite an ongoing debate on this causal approach[1], the automated causal inference based on observed data has become a powerful instrument to assess causal relations empirically.

Recently, these graphical models have found their way into the literature on time series analysis and econometrics. Eichler (2007) gives a graphical presentation of the Granger causality among multivariate time series. Some pioneering works of graphical causal models in econometrics can be found in Glymour and Spirtes (1988). Hoover (2005) sketches the application of the graphical causal approach to identification of structural VAR models. Swanson and Granger (1997) apply a similar concept to identify the causal chain in VAR residuals. Demiralp and Hoover (2004) apply the graphical causal method to VAR residuals to infer the causal orders in the money demand and the monetary transmission mechanism.

Along this line of research we extend the application of the graphical causal approach to VAR residuals to the time series themselves and infer the causal orders in the multivariate time series. As a byproduct, the causal orders in the residuals are determined by the causal orders in the time series. Concretely, we view N time series with T observations as realizations of a set of NT random variables and embed these NT random variables into a directed acyclic graphical (DAG) model with NT nodes. The aim of this paper is to develop an effective method to infer the causal relations among these NT random variables.

The paper is organized as follows.

In Section 2 we review shortly the basic idea and features of the theory of inferred causation. Then we embed multivariate time series into a DAG model to define a time series causal model(TSCM). We formulate assumptions under which a TSCM can be represented through a partial DAG and hence becomes statistically assessable. We discuss the relation between structural vector autoregression and TSCM and derive the Granger causality in a TSCM. In Section 3 we present a learning algorithm to infer the causal relations among time series variables and document simulation results to assess the performance of the learning algorithm. In Section 4 we apply TSCM to analyze the wage-price spiral in the Australian economy. The last section concludes.

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

2 Time Series Causal Model

2.1 Theory of Inferred Causation

2.1.1 Causal Models

The basic idea of theory of inferred causation is to present a causal structure among variables in an acyclic directed graph (DAG) called a causal graph in which arrows indicate causal orders. Based on a generally established relationship between topologies of causal graphs and conditional independencies among variables in the graphs, sample information on conditional independencies of a set of variables is used to infer the topology of the data-generating causal graph and the direction of arrows in the graph. In this way the causal structure among variables can be inferred from empirical observations of the variables. Pearl (2000) gives a systematic account of the theory of inferred causation and Spirtes et al. (2000) discuss in detail the techniques and algorithms used to uncover the data-generating DAGs.

Formally the theory of inferred causation is built on a fundamental assumption on the cause-effect relations as given in following definitions in Pearl (2000).

Definition 2.1 (Causal Structure in Pearl (2000) p.44) *A causal structure of a set of variables V is a directed acyclic graph(DAG) in which each note corresponds to a distinct element of V, and each link represents direct funcitonal relationship among the corresponding variables.*

Definition 2.2 (Causal Model in Pearl (2000) p.44) *A causal model is a pair $M =< D, \Theta >$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $P(u_i)$ to each u_i, where PA_i are parents*[2] *of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$ independently of all other u.*

Probability measures compatible with D are called to satisfy the causal Markov condition. The causal Markov condition implies that conditioning on $PA(X_i)$, X_i is independent of all its nondescendants. In particular it implies that the disturbance U_i are independent from other Us. In addition to the causal Markov condition, the minimality of the causal structure[3] D, and the stability of the distribution[4] 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[5]. If there is an arrow from X_i to X_j we say X_i is a direct cause of X_j . If there is sequence of arrows, all pointing in one direction from X_i to X_j, we say X_i is an indirect cause of X_j.

In Fig. 1, X_3 is called a predecessor of X_5, because there is a directed path from X_3 to X_5. X_2 is called a parent of X_1 and X_3, because X_2 is a direct predecessor of X_1 and X_3. The two arrows $X_1 \to X_5$ $X_3 \to X_5$ constitute a v-structure, because the two arrows are heading at X_5 and their ends are not connected.

[2]Parents are direct predecessors.

[3]See Definition 5 in Pearl and Verma (1991) and Definition 2.3.4 in Pearl (2000) p.46.

[4]see Pearl (2000) p.48 and p. 61. and Spirtes et al. (2000) p. 29 ff.

[5]It is still an ongoing debate whether causality can be formulated in such assumptions. See Cartwright (2001), Pearl (2000) p. 41, Spirtes et al. (2000) p. 105. Freedman and Humphreys

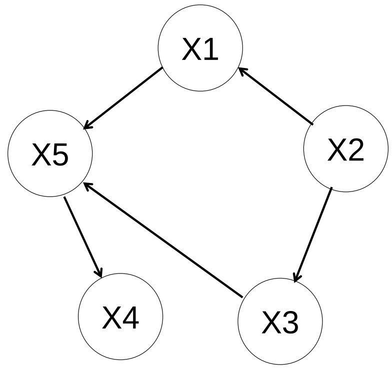

Figure 1: Influence Diagram

A compatible distribution of a DAG can be factored into the conditional distributions according to the DAG. For example we know that for the DAG in Fig.1 the joint distribution can be calculated as follows

$$f(x_{1t}, x_{2t}, x_{3t}, x_{4t}, x_{5t}) = f(x_{4t}|x_{5t})f(x_{5t}|x_{1t}, x_{3t})f(x_{3t}|x_{2t})f(x_{1t}|x_{2t})f(x_{2t}).$$

x_{it} is a realization of X_{it}. The DAG in Fig. 1 implies following conditional independencies: 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}. These conditional independencies can be used to infer the arrows in the DAG in Fig. 1.

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 data generating causal structure using observed data, and then to interpret the directed edges in the DAG as cause-effect relations. Identifying the data generating DAG from the patterns of conditional independencies and dependencies is one of the main research activities in the area of inferred causation.

2.1.2 Observational Equivalence and Inferrable Causation

If data are generated from a causal model, can statistical procedure always uniquely recover the data-generating causal structure? The answer to this question leads to the problem of observational equivalence of a causal model. Observationally equivalent models will generate data with identical statistical properties. Therefore, statistical method can identify only the underlying DAGs up to the observationally equivalent classes. For the observational equivalence of causal models we quote the result in Pearl (2000) p.19.

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

Because statistical method cannot differ the observationally equivalent DAG models from each other, not every causal direction in a DAG can be inferred. Only

(1998) for more discussion. Spirtes et al. (2000) took an axiomatic approach to pave the logical basis for the method of inferred causation.

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. We call these causal directions the *inferrable causal directions*. If a data generating DAG has observationally equivalent models, the directions of some arrows in the DAG cannot be uniquely inferred from the data. Hence, the existence of observational equivalence places a limit on the ability of statistical method in inferring causal directions.

Given a set of data generated from a causal model, a statistical procedure can principally identify all the conditional independencies. However, the statistical procedure cannot tell whether this kind of independencies are due to the absence of some edges in the DAG of the causal model or due to the particularly chosen parameter values of the causal model such that these edges in this case imply the conditional independencies. To rule out this ambiguity, Pearl (2000) assumes that all the identified conditional independencies are due to absence 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.

2.1.3 Search Algorithms

To infer the data generating causal graph from sample information is call learning of the graph in the literature. There are basically three kinds of solutions to this learning problem. The first solution is based on sequential tests of partial correlation coefficients. The tests run from the lower order partial correlation coefficient in unconstrained models to the higher order partial correlation coefficients. Hoover (2005) gives a very intuitive description of this procedure. Spirtes et al. (2000) provide an elaborated discussion about this kind of algorithms[6]. A simple version of the most popularly used PC algorithm is given as follows.[7]

PC Algorithm

Input: Observations of a set of variables X generated from a DAG model.
Output: a pattern (DAG) compatible with the data generating DAG.

- Start with a full undirected graph. For each pair of variables $(X_i,X_j) \in X$, search a subset $S_{ij} \in X/\{X_i, X_j\}$ such that $(X_i \perp X_j | S_{ij})$[8] holds, then delete the edge between X_i and X_j.

- 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 it 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

[6] *PC* algorithms named according to its inventors Peter Spirtes and Clerk Scheines is the most popular algorithm in uncovering causal graphs. See http://www.phil.cmu.edu/projects/tetrad/ for more details and software for this algorithm.

[7] For our presentation purpose, we give here a simplified version of PC algorithm. For more sophisticated version of *PC* algorithm see Spirtes et al. (2000) p. 89.

[8] $(X_i \perp X_j | S_{ij})$ means, conditioning on S_{ij}, X_i is independent from X_j

create a new v structure; and (ii) the orientation should not create a directed cycle.

Since the tests in the PC algorithm are consistent, with increasing number of observations and a significance level approaching zero the probability to identify the edges correctly based on the tests will converge to one. This fact is summarized in the following proposition.

Proposition 2.3 *Under the assumption of faithfulness, the PC-algorithm can consistently identify the inferrable causal directions, i.e. for $T \to \infty$ the probability of recovering the inferrable causal structure of the data generating causal model converges to one.*

Proof: (See Robins, Scheines, Sprites, and Wasserman (2003)) □

This Proposition says in particular that if the data generating causal model has no observational equivalence, the PC-algorithm will uniquely identify the causal structure consistently. If the data generating causal model has observational equivalence, the PC-algorithm will uniquely identify the observational equivalent class.

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 called the score of a graph, and in the choice of search strategy. Because the search problem is NP-hard[9] heuristic search algorithms such as greedy search, greedy search with restarts, best-fit search, and Monte-Carlo method are used[10]. 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 graph is evaluated according to information criteria such as AIC or BIC. The search algorithms are similar as those in the Bayesian approach, such as greedy search, and greedy search with restarts. A simple version of the greedy search algorithm is given as follows.

Greedy Search Algorithm:
Input: Observations of a set of variables X generated from a DAG model.
Output: a pattern (DAG) compatible with the data generating DAG.

- Step 1 Start with a DAG A_o.
- Step 2 Calculate the score of the DAG according to BIC/AIC/likelihood criterion.
- Step 3 Generate the local neighbour DAGs by either adding, removing or reversing an edge of the network A_o.
- Step 4 Calculate the scores for the local neighbour DAGs. 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 with A_n. If the highest scores is less than that of A_o, stop and output A_o.

[9]See Heckerman (1995) for details.

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

It is to note that a causal model is a statistical model. If the score used in the greedy search algorithm is a consistent model selection criterion such as BIC, the greedy search algorithm will consistently recover the inferable causal directions, presuming that the search space covers the true DAG.

2.2 Time Series Causal Models

2.2.1 DAGs and Recursive Structural Models

It can be shown that if an n-dimensional variable X is jointly normally distributed, a linear causal model of X is equivalent to a linear recursive structural equation model (SEM)(See Pearl (2000), p. 141.). The function that associates a variable with its parents can be written as follows.

$$x_j = \sum_{k=1}^{j-1} a_{jk} x_k + u_j \qquad \text{for} j = 1, 2, ...n, \tag{2.1}$$

where u_j are independently normally distributed. We call (2.1) the structural equation of the linear causal model. We summarize this fact in the following proposition.

Proposition 2.2 *If a set of variables X are jointly normal $X \sim N(0; \Sigma)$, a linear causal model for X can be equivalently formulated as a linear recursive structural equation model (SEM) that is represented by a lower triangular coefficient matrix A with ones on the principal diagonal. Any nonzero element in this coefficient matrix, say α_{jk} corresponds to a directed edge from variable k to variable j.*

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

where A is the triangular decomposition matrix of Σ with $A\Sigma A' = \Lambda$ and Λ is a diagonal matrix.

Proof: See Pearl (2000) P. 141-142.

2.2.2 Time Series Causal Models

The linear causal model presented in the last subsection is applicable to independent data. Economic time series are, however, dependent data. Nevertheless, we can view N time series with T observations as realization of NT random variables. We can embed these NT random variables into a large recursive structural equations model. Under the assumption that the elements of the multivariate time series X_{it}, $i = 1, 2, ..., N$ and $t = 1, 2, ...T$ are jointly normal, then following Proposition 2.2 a causal model for the multivariate time series is a linear recursive structural model in all the NT components.

Since temporal information provides a nature causal order, the recursive structural model must follow the temporal order. Hence, we can write the recursive system as follows.

$$\begin{pmatrix} A_{11} & 0 & \dots & 0 \\ A_{21} & A_{22} & & 0 \\ \vdots & & \ddots & \vdots \\ A_{T1} & A_{T2} & \dots & A_{TT} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}, \tag{2.2}$$

where $\epsilon_t \sim N(0, D)$ is a vector of independent residuals and D is a diagonal matrix, ϵ_t and $\epsilon_{t-\tau}$ are independent, and $X_t = (X_{1t}, X_{2t}, ..., X_{Nt})'$ for $t = 1, 2, ...T$ is the random vector at time t[11].

Because we only have one observation at each time point, the recursive system (2.2) contains too many parameters to be analyzed statistically. Therefore we need to impose reasonable constraints on the parameters of the system to make the system statistically assessable. Following Chen and Hsiao (2007) beside the temporal causal constraint, two reasonable assumptions are the time-invariant causal structure constraint that the causal structure between variables at time points t and s is the same as the causal structure between variables at time points $t + \tau$ and $s + \tau$, and the time-finite causal influence constraint that X_t may have a causal influence on $X_{t+\tau}$ only when $\tau \leq p$, where $p < \infty$ is a given positive integer. Under the assumptions of the temporal causal constraint, the time-invariant causal structure constraint and the time-finite causal influence constraint, the linear recursive system (2.2) with $p = 2$ can be written as follows.

$$\begin{pmatrix} A_0 & 0 & \dots & & \dots & 0 \\ A_1 & A_0 & 0 & & \dots & 0 \\ A_2 & A_1 & A_0 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & A_2 & A_1 & A_0 & 0 \\ 0 & \dots & 0 & A_2 & A_1 & A_0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{T-1} \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{T-1} \\ \epsilon_T \end{pmatrix}. \tag{2.3}$$

The parameter matrices $A_1, A_2, ...A_p$ at t-th row present the causal influence of $X_{t-1}, ...X_{t-p}$ on X_t and A_0 is the contemporaneous causal influence among the elements of X_t. The time-finite constraint implies that in each row all the parameter sub-matrices left to A_p are zero. We call the causal model in (2.3) a time series causal model (TSCM).

Since the coefficient matrix in (2.3) is a lower triangular matrix, A_0 must be a lower triangular matrix too. Equation (2.3) can be reformulated as follows[12].

$$A_0 X_t + A_1 X_{t-1} + ...A_p X_{t-p} = \epsilon_t, \qquad \text{for } t = p + 1, p + 2, ..., T. \tag{2.4}$$

Corresponding to the TSCM in (2.4) we can represent the DAG for a TSCM through a partial DAG, namely only through $(p + 1)N$ nodes representing $X_t, X_{t-1}, ..., X_{t-p}$ and the arrows heading at the elements of X_t (see Fig. 2). This implies that instead of a DAG with TN nodes we need now only to consider a partial DAG with $(p+1)N$ nodes. In the following subsections we are going to discuss the observational equivalence in a partial DAG and the learning of a partial DAG.

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

[12]We take the initial value as given.

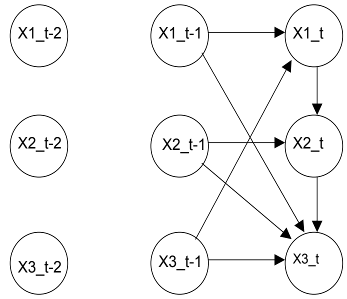

Figure 2: Partial DAG of a TSCM

2.2.3 Observational Equivalence in TSCMs

Since TSCMs are specifically restricted DAG models, the result of Proposition 2.1 still holds. However, since we present TSCMs in partial DAGs we need to reformulate the proposition in terms of partial DAGs. Because the arrow direction from X_{t-i} into X_t is fixed, we need only to consider the observational equivalence due to direction changes of arrows between the elements of X_t. For an arrow connecting two elements of X_t, say X_{it} and X_{jt}, if X_{it} and X_{jt} have different parents, then a change of the arrow direction will lead to a new v structure. Based on this fact we can formulate the condition of the existence of observational equivalence.

Proposition 2.4 *A partial DAG has an observationally equivalent model if there are some arrows between elements of X_t that satisfy the following two conditions*

- *the lagged parents of the connected elements of X_t are the same, and*
- *the change of the arrow directions will not lead to a new v-structure or a cycle in the partial DAG.*

Corollary 2.5 *If in a partial DAG all the elements of X_t have different lagged parents, the partial DAG does not have an observationally equivalent model.*

2.2.4 TSCMs and VAR Models

TSCMs are motivated by causal modeling to infer causal relations in time series data. Formally, they are linear relations between time series variables at present period and the time series variables in the past periods as well as at present period. Therefore, there is an intimate relation between TSCMs and the popular vector autoregressive models in time series econometrics. We summarize this relation in the following two propositions.

Proposition 2.6
A TSCM is a restricted structural VAR model identified by the inferred causal relations among $\{X_t\}_{t=1}^T$, and hence it corresponds to a restricted VAR model.

Proof: Since a full DAG does not contain any inferable causal relations, a sensible TSCM will have some null restrictions in A_0, A_1,...,A_p therefore they corresponds to a restricted SVAR model in the causal order of X_t. Since A_0, A_1,...,A_p are subject to some restrictions, the coefficients of the corresponding reduced form VAR:

$$\begin{aligned} X_t &= -A_0^{-1}A_1X_{t-1} - A_0^{-1}A_2X_{t-2}+, ..., +A_0^{-1}A_pX_{t-p} + A_0^{-1}\epsilon_t \\ &= \Pi_1X_{t-1} + \Pi_2X_{t-2} + ... + \Pi_pX_{t-p} + e_t \end{aligned} \tag{2.5}$$

are also subjected to some restrictions. □

Proposition 2.7
An unconstrained VAR model corresponds to a full partial DAG such that the TSCM does not contains any inferrable causal relations except the temporal causal orders.

Proof: An unconstrained VAR corresponds to an unconstrained SVAR in an arbitrary order of the elements of X_t by using Cholesky decomposition of the covariance matrix of the residuals. A unconstrained SVAR corresponds to a full partial DAG in which every node of $X_{i,t-s}$ $(s \leq p)$ is connected to all N elements in X_t and the N nodes of X_t constitute a full DAG among themselves. In this case all elements of X_t have same lagged parents and there is no v structure consisting of arrows connecting two elements in X_t. Therefore, the partial DAG does not imply any inferable causal relations except the temporal causal orders. □

2.2.5 Granger Causality in TSCM

In time series analysis one often used concept is Granger causality. Given a TSCM we can derive the Granger causality among the time series variables in the TSCM. Generally, Granger causality and the graphic causal models are two different concepts: while the Granger causality concerns the prediction power of one time series for the another, a TSCM concerns the causal relation among time series variables at each time points. The following proposition gives how the Granger causality among the time series in a TSCM can be derived.

Proposition 2.8
Let $X_{i,t}$ and $X_{j,t}$ be two time series variables in a TSCM. $X_{j,t}$ is a Granger cause of $X_{i,t}$ given other variables in the TSCM if and only if there is a directed path from some $X_{j,t-s}$ to $X_{i,t}$ for $s > 0$ in the partial DAG of the TSCM.

Proof (See Appendix).

This Proposition provides a causal insight into the multivariate Granger causality between two time series. If a lagged $X_{j,t}$ has a causal influence on $X_{i,t}$ representing by a directed path from the lagged $X_{j,t}$ to $X_{i,t}$, then the lagged $X_{j,t}$ will contain a unique information about $X_{i,t}$ that is not included in the past of $X_{i,t}$ and the past of other relevant variables. If the other relevant variables include all carefully chosen explanatory variables, the unique information embodied in lagged $X_{j,t}$ justifies to qualify the prediction ability as "causality". This proposition says also that if there no directed or indirect causal influence from $X_{j,t-s}$ on $X_{i,t}$ in addition to the causal influences from other $X_{k,t-s}$ $(k = 1, 2, j-1, j+1, ...N)$ on $X_{i,t}$, $X_{j,t}$ is not a Granger cause of $X_{i,t}$.

3 Learning TSCM

3.1 Learning TSCM

For a TSCM we need only to learn a partial DAG with $(p+1)N$ nodes instead of the complete DAG with TN nodes. Given that we want to learn a partial DAG consisting of all arrows into the nodes at time point t, what is the pupulation/sample information that allows a correct inference on arrows in the partial DAG? The following proposition answers this question.

Lemma 3.1 *Given the assumption of a causal model, an information set (joint distribution) containing a node and a set of variables including its parents is sufficient for PC algorithm to connect the node to all its parents and exclude all its non-descendants from connecting to it.*

Proof:

According to the causal Markov assumption, conditional on the parents of a variable, this variable is independent from all its non-descendants. Since the parents of the concerning variable are all included in the information set, PC algorithm will all edges between the variable and its non-descendants.

On the other hand, for the concerning variable and one of its parents, there is no subset of the total NT variables excluding these two variables, such that conditioning on this subset the concerning variable and the parent are independent. Since the information set is a subset of the total NT variables, it follows that there is no subset of the information set excluding these two variables, such that conditioning on this subset the concerning variable and the parent are independent. Therefore, no edge between the variable and its parents will be missing. □

Proposition 3.2 *To learn the partial DAG with arrows into X_t the information set including $X_t, X_{t-1}, ..., X_{t-p}$ is sufficient.*

Proof:

Since the information set contains all parents of X_t, Lemma 3.1 above establishes that all arrows from X_{t-i} into X_t will be correctly inferred. We need only to make sure that the arrows between the elements of X_t are also inferred correctly. Considering two nodes X_{it} and X_{jt}, if one is a parent of the other, an edge will be inferred according to Lemma 3.1. If there is no parent-child relationship between the two variables, then one of them must be non-descendant of the other, according to Lemma 3.1 there will be no edge between them.

Concerning the direction of the arrows, following the assumption of temporal causal constraint, the arrows always go from X_{t-i} to X_t. Among the the edges connecting elements of X_t, the rule of orientation in PC algorithm implies that the orientation based on the information set is the same as the orientation based on the total variables. □

It is to note that Proposition 3.2 says that applying PC algorithm based on the information set containing $(X_t, X_{t-1}, ..., X_{t-p})$ will only give correct arrows into X_t. The arrows and edges among X_{t-i} may be incorrect. However, correct inference of the arrows heading into X_t is sufficient to generate a correct partial DAG.

According to Proposition 3.2 we have the following algorithm to learn the partial DAG for a TSCM.

A Modified PC Algorithm for a Partial DAG

Input: Observations of a set of time series variables X generated from a TSCM. Output: a partial DAG compatible with the data generating DAG.

- step 1: Choose a reasonable $\hat{p}$
- step 2: Calculate the correlation matrix of $\Sigma = \text{corr}(X_t, X_{t-1}, ...X_{t-\hat{p}})$
- step 3: Using Σ as input to obtain a DAG for $(X_t, X_{t-1}, ...X_{t-\hat{p}})$
- step 4: Delete all arrows and edges that do not connect at least one element of X_t
- step 5: Orient all edges between X_{t-i} and X_t with arrowheads at X_t.
- step 6: Orient all edges between elements of X_t using the rules in PC algorithm.

Remarks The choice of $\hat{p}$ determines the lag length of the TSCM. If a chosen $\hat{p}$ is smaller then the true p, some direct parents of X_t will not be in the information set, inference on the edges connecting X_t will be incorrect. If $\hat{p}$ is larger than p all parents of X_t are included in the information set. Therefore, the inference of edges connecting X_t is correct. If in the output partial DAG no arrows go from $X_{t-\hat{p}}$ to X_t, this indicates that the choice of $\hat{p}$ is large enough. However, a large $\hat{p}$ will lead to a larger graph with more nodes and will, hence, reduce the power of tests in finite samples.

For a DAG model, evaluating graph scores is an alternative way to uncover the data generating DAG model. For a partial DAG, what is the proper score of the graph? We know that a partial DAG corresponds to a SVAR model as in (2.4), it is natural to use the likelihood of (2.4) to evaluate the model. Since unconstrained model (2.4) will always have higher likelihood than a constrained model (2.4), a proper score can be an information criterion that adds a penalty term to the likelihood due to the dimension of the model. For a partial DAG of X_t we can define the BIC criterion as follows

$$BIC = \sum_{t=1}^{T} \log L(A_0, A_1, ..., A_p; X_t | X_{t-1}, ..., X_{t-p}) - (|E| + |V|) \log(T),$$

where $|E|$ is number arrows heading at X_t in the partial DAG and $|V|$ is the number of elements in X_t. The sum of $(|E| + |V|)$ is just the number of free varying parameters of the TSCM under consideration. This BIC criterion is a sum of the log likelihood function value and the number of parameters of the model times a penalty factor $\log(T)$. As the penalty factor satisfies the condition (1) $\log(T) \to \infty$ as $T \to \infty$, (2)$\frac{\log(T)}{T} \to 0$ as $T \to \infty$, and the log likelihood function grows at rate T, the BIC criterion is a consistent model selection criterion for TSCMs. We summarize this fact in the following Proposition.

Proposition 3.3 *Under the assumption of TSCM, the BIC criterion is a consistent score, such that the probability of identifying the true model will converge to 1 as $T \to \infty$, presuming that the search space covers the true model.*

As in the case of a DAG, we can also apply a greedy search algorithm to learn a partial DAG. Problem with greedy search algorithms is that it finds only a local optimum. A good starting graph is crucial for a good performance of this search method. Because *PC* algorithm give a consistent partial DAG and *PC* algorithm converges very fast, its output provides a good initial graph for the greedy search algorithm. We will show in next subsection that a combination of the *PC* algorithm and greedy search will greatly improve the performance of the causal learning algorithm.

Remarks It is to note that the learning algorithm presented above will infer a causal structure if the data are generated from a TSCM with inferrable causal relations. If data are generated without any causal orders, the learning algorithm will give a DAG without any inferable causal directions, such as a full partial DAG. In this sense, the learning algorithm follows automatically the general to specific modeling strategy[13]. In the stage with *PC* algorithm to find a proper starting graph, it goes from a more general model, i.e. a full partial DAG with a maximum lag and test down to a more restrictive model with less arrows and less lags. In the greedy search stage, the selected model is compared to its local alternative to obtain a better model according to their respective scores.

3.2 Simulation Studies

The results of the learning procedure presented in the last section are asymptotically valid. For empirical applications, small sample properties of the procedure are more relevant. In this subsection we conduct a simulation study to assess the performance of the learning procedure in small sample situations.

The data generating process in the simulation study is as follows:

$$A_0X_t + A_1X_t = u_t. \tag{3.6}$$

In this data generating process we consider only one lag. It is, however, less restricted as it appears, because TSCMs with more lags can be equivalently represented as a TSCM of a higher dimension with only one lag (See Hamilton (1994) p. 7 for more details.).

A_0 is set to be a lower triangular with ones on the principle diagonal. Other non-zero elements of A_0 are random numbers from a uniform distribution over $[1, 2]$. The nonzero elements of A_1 are random numbers from a uniform distribution over [0.4,0.9]. The zero elements in A_0 and A_1 are chosen randomly. The parameters in A_0 and A_1 are chosen under the restriction that the time series are stationary. The dimensions of X_t are chosen to be 3, 4 and 5, and the number of observations are 100, 200 and 400, which include the most often encountered application cases. u_t is iid normally distributed with variance one.

We summarize the simulation results as follows

- In the designed setting, *PC* algorithm performs poorly in recovering a complete data-generating DAG. In all simulation runs the percentage of correctly recovered graphs is zero (See the column under the header *PC*% in Table 1.). However, the percentage of correctly recovered arrows are much higher. It varies from 81% to 84%. The relative high frequencies implies that the

[13] See Hoover (2005) for a more elaborated discussion on this point.

T	D	p	GS_PC	IC	PC	GS	PC_a	GS_PC_a
100	3	2	83.6	4.7	0	81.4	81.4	98.3
200	3	2	99.4	1.2	0	81.6	81.6	99.9
400	3	2	99.7	0.0	0	81.4	81.4	99.9
100	3	3	96.7	4.1	0	82.6	82.7	99.9
200	3	3	99.5	1.8	0	82.4	82.4	99.6
400	3	3	99.0	0.8	0	81.4	82.6	99.9
100	4	2	89.5	6.8	0	75.5	81.5	99.3
200	4	2	96.6	1.1	0	82.5	84.3	99.6
400	4	2	97.9	0.0	0	81.5	81.2	99.0
100	4	3	81.5	8.1	0	77.5	82.2	99.4
200	4	2	93.7	2.3	0	80.5	88.5	98.8
400	4	3	96.7	0.0	0	76.1	87.8	99.3
100	5	2	83.0	8.1	0	70.5	83.7	98.2
200	5	2	93.5	5.3	0	78.5	83.8	99.4
400	5	2	96.2	0.0	0	76.5	83.7	98.3
100	5	3	82.2	9.1	0	71.5	83.4	99.0
200	5	3	92.1	6.3	0	75.5	84.1	99.3
400	5	3	95.4	0.0	0	79.5	83.9	99.1

Each row in the table records a simulation result of 1000 runs. First column under the header T gives the number of observations used in each simulation. D is the dimension of X_t, p is the lag length used in the learning procedures. The numbers under the header GS_PC record the frequencies of correctly recovered causal structures by using a greedy search with PC output as starting graphs. The column under the header IC record the percentages where the score value of the true model is not the maximal score. The column of PC records the percentages of correctly identified causal structures using PC algorithm. The column of GS records the percentages of correctly identified causal structures using the greedy search algorithm with a random starting graph. PC_a gives the percentage of correctly identified arrows in the graphs using PC algorithm. GS_PC_a gives the percentage of correctly identified arrows in the graphs using the greedy search algorithm with PC output as starting graphs.

Table 1: Simulation results of recovering A_0 and A_1 in equation (3.6)

.

graphs identified by the PC algorithm are very similar to the corresponding true graphs, but with one or two wrong arrows. This suggests that the output of PC can be a good starting graph for greedy search algorithm.

- The greedy search algorithm with a random staring graph performs better than the PC algorithm but its performance is not very satisfactory (See the column under the header GS in Table 1.). The percentage of correctly recovered graphs ranges from 70% to 82%. Because of the nature of a local search, the performance of the greedy search algorithm depends crucially on the starting graph.

- Using PC output as an initial graph for the greedy search algorithm improves the performance of the search algorithm greatly. Overall the results are satisfactory. The percentages of correctly identified data-generating DAG varies from 83% to 98% (See the column under the header GS_PC in Table 1.).

With increasing number of observations the percentage of correctly identified data generating DAG is getting higher. The percentage of correctly identified arrows in each simulation runs are over 95%.

4 An Application to Wage - Price Spiral Dynamic System

In this section we apply the TSCM developed in previous sections to analysis the causal structure in the wage-price spiral in the Australian economy in order to answer the question whether wage inflation causes price inflation or the other way around. A bivariate Granger-causality test for the two time series: dp_t the price inflation and dw_t the wage inflation gives the following result.

```
         F-statistic    p-value
DW -> DP    3.254229 0.01449963
DP -> DW    3.158491 0.01682668
```

The mutual Granger Causality is often seen as an evidence that supports wage-price spiral hypothesis, which suggests that rising wages increase income, thus increase the demand for goods and cause prices to rise. Rising prices cause demand for higher wages, that leads to higher production costs and further upward pressure on prices. This is a reason why it is called wage-price spiral. However, the mutual Granger causality does not necessarily implies that they are mutual cause to each other.

To investigate the mechanism behind this mutual temporal dependence, we adopt the theoretical framework as set out in Flaschel and Krolzig (2003) and Chen and Flaschel (2006), in which two Phillips curves, one for price inflation and one for wage inflation are used to describe the dynamic wage-price spiral. The theoretical formulation of the Phillips curves are as follows.

$$dw = \beta_{w1}(V^l - \bar{V}^l) + \kappa_w dp + (1-\kappa_w)\pi^m + \beta_{w2}dz \quad (4.7)$$
$$dp = \beta_{p1}(V^c - \bar{V}^c) + \kappa_p dw + (1-\kappa_p)\pi^m + \beta_{p2}dz \quad (4.8)$$

In these symmetrically formulated two Phillips curve equations, we can describe wage and price dynamics separately from each other. Both variables react to their own measure of demand pressure: namely $V^l - \bar{V}^l$ and $V^c - \bar{V}^c$, in the market for labor and for goods, respectively. We denote by V^l the rate of labour utilization on the labor market and by $\bar{V}^l$ the NAIRU-level of this rate, and similarly by V^c the rate of capacity utilization of the capital stock and $\bar{V}^c$ the normal rate of capacity utilization of firms. These demand pressures are both augmented by a weighted average of cost-pressure terms. Cost pressure perceived by workers is a weighted average of the currently evolving rate of price inflation dp and the expected price inflation, π^m. Similarly, cost pressure perceived by firms is given by a weighted average of the currently evolving rate of wage inflation, dw and again the measure of expected inflation. Further the Phillips curves are augmented by changes of labor productivity dz that impacts positively on the wage inflation and negatively on the price inflation (see Flaschel and Krolzig (2003) for more details of theoretical arguments on this type of two Phillips curves.)

The two Phillips curves present a theoretical hypothesis how the wage-price spiral is interacting. The objective of our empirical analysis is to infer the causal relations among the 6 variables involved in these two Phillips curves in order to investigate in how far the causal relations implied by the observed data can support the hypothetical formulation of wage price spiral as given in (4.7) and (4.8), and to investigate the mechanism behind the mutual Granger causality between the wage inflation and the price inflation.

The empirical data for the relevant variables are taken from Australian Bureau of Statistics[14]. The data shown below are quarterly, seasonally adjusted, annualized where necessary. The data used in this investigation are from 1978:3 to 2009:2, which correspond to the longest commonly available time series for the set of variables used in the investigation.

Variable	Transformation	Description
e	$100 - URATE$	URATE: Unemployment Rate(%) e: Employment Rates
u	$\frac{GDP}{GDP\_HPtrend}100$	GDP: Real Gross Domestic Product Chain volume measures. DGP_HPtrend: the trend component of HP filter applied to GDP. u: Capacity utilization rate, ratio
dw	$\frac{AWE-AWE(-1)}{AWE(-1)}400$	AWE: Average Weekly Earnings, dw : wage inflation, annualized
dp	$\frac{CPI-CPI(-1)}{CPI(-1)}400$	CPI: Consumer price index, all groups, Index 1990 = 100 dp : price inflation, annualized
z	$\frac{GDP}{HOURS}$	HOURS: Total (Actual hours worked) z: labor productivity
dz	$\frac{z-z(-1)}{z(-1)}400$	dz: change of labor productivity, annualized
π^m :	CIE	Consumer inflation expectation (%), survey data, Westpac-Melbourne Institute Consumer Survey.

Table 2: Raw data used for empirical investigation of the wage-price spiral

We construct a TSCM consisting of six time series variables $(dp, dw, \pi^m, e, u, dz)$[15]. Through a series of unit root tests dp, dw, π^m, e, u, dz are confirmed to be stationary,

[14]See the web site for more details. http://http://www.abs.gov.au/

[15]We correct the data of dp with a dummy variable d_GST, to take into account of the impact of the introduction of the good and service tax (GST) on prices in the third quarter 2000.

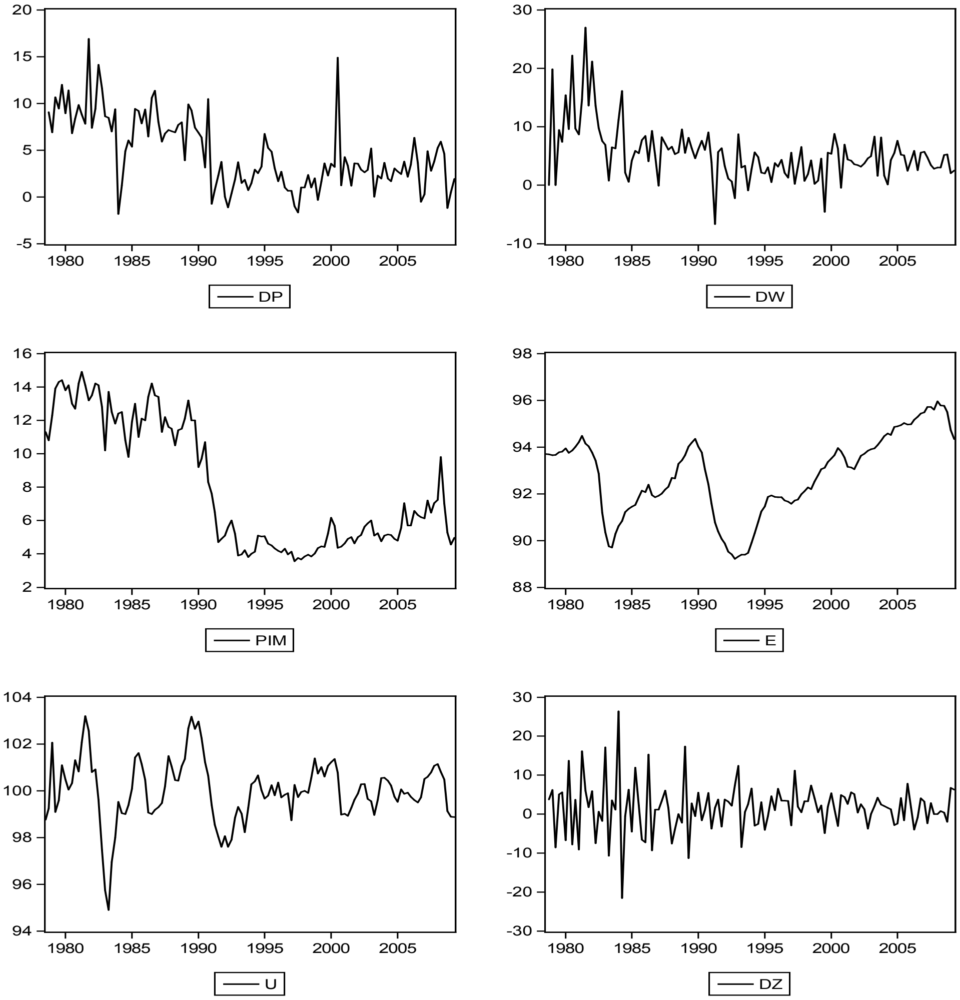

Figure 3: Data for the analysis of wage-price spiral

where the unit test for π^m is run after controlling for a structural break in 1991:2.

We take $p = 4$ to derive the partial DAG for the TSCM. The choice of $p = 4$ is to make sure that the lag length is chosen long enough to avoid bias in the specification. If the lag length of the true TSCM is less than 4, the estimated partial DAG will not include any arrows from X_{t-4} to X_t. Figure 4 shows the output of the procedure.

In the partial DAG in Figure 4 there are no arrows from $(dp_{t-4}, dw_{t-4}, \pi^m_{t-4}, e_{t-4}, u_{t-4}, dz_{t-4})$ into $(dp_t, dw_t, \pi^m_t, e_t, u_t, dz_t)$. This implies that the TSCM has a lag length of 3. One important feature of this partial DAG is that $(dp_t, dw_t, \pi^m_t, e_t, u_t, dz_t)$ have different lagged parents. Following Proposition 2.4 the inferred DAG does not have any observationally equivalent models, i.e. all arrow directions in the partial DAG in Fig. 4 are uniquely determined by the data.

The partial DAG says that dp_t is influenced by π^m_t and u_{t-3}; and dw_t is influenced by π^m_{t-1}, e_t and dz_{t-1}. But dp_t, dw_t and their lags don't influence other variables: dz_t, π^m_t, e_t and u_t. In other words the latter four variables are determinants of the price inflation and wage inflation. Further, the causal structure of the partial DAG explains that the mutual bivariate Granger causality between dp_t and dw_t is the effect caused by common cause variables: π^m has a direct influence on both dp and

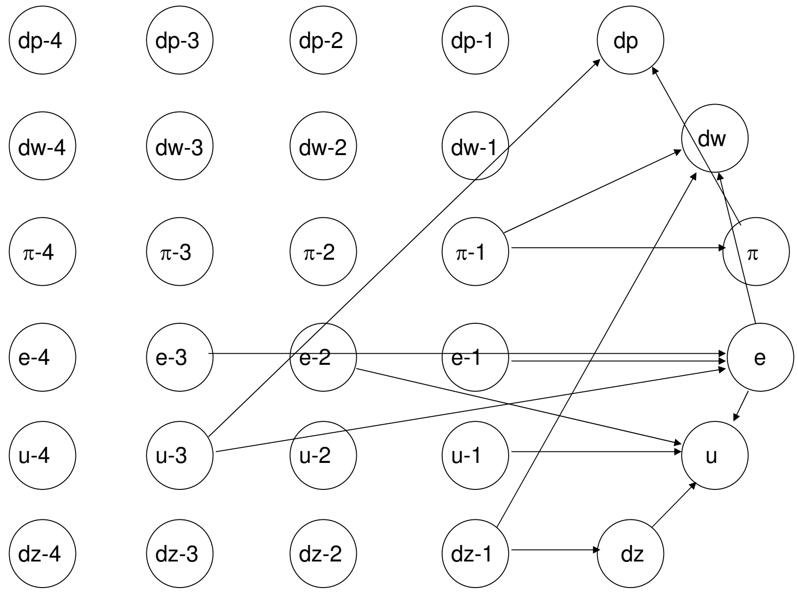

Figure 4: Partial DAG of the Wage Price Spiral

dw; dz has a direct influence on dw and an indirect influence via u on dp; e has similarly a direct influence on dw and an indirect influence via u on dp; and dz has a direct influence on dw and an indirect influence on dp.

The linear causal equations derived from this TSCM are

$$dp_t = \underset{14.69}{0.76}\pi_t^m + \underset{3.50}{0.52}u_{t-3} - \underset{-3.61}{53.03} + \epsilon_{pt} \tag{4.9}$$

$$dw_t = \underset{8.34}{0.72}\pi_{t-1}^m + \underset{2.59}{0.48}e_t + \underset{3.94}{0.21}dz_{t-1} - \underset{-2.63}{45.28} + \epsilon_{wt} \tag{4.10}$$

$$\pi_t^m = \underset{40.92}{0.96}\pi_{t-1}^m + \underset{-1.18}{0.23} + \epsilon_{\pi^m t} \tag{4.11}$$

$$e_t = -\underset{-4.57}{0.08}u_{t-3} - \underset{-9.81}{0.36}e_{t-3} + \underset{38.13}{1.35}e_{t-1} + \underset{5.72}{9.06} + \epsilon_{et} \tag{4.12}$$

$$u_t = -\underset{-6.88}{0.67}e_{t-2} + \underset{7.00}{0.72}e_t + \underset{14.17}{0.73}u_{t-1} + \underset{3.83}{0.03}dz_t - \underset{4.95}{22.41} + \epsilon_{ut} \tag{4.13}$$

$$dz_t = -\underset{-4.81}{0.40}dz_{t-1} + \underset{4.3}{2.23} + \epsilon_{zt}. \tag{4.14}$$

The first two equations (4.9) and (4.10) are the inferred structural Phillips curves for the price inflation and the wage inflation respectively. Unlike most empirically estimated Phillips curves, these two structural equations have the following features: (1) these two Phillips curve equations possess a causal interpretation: the right hand side variables have causal influence on the left-hand side variables; (2) the causal relations are not imposed on the variables a priori, but derived from the observed data using the theory of inferred causation: they are obtained through a data-driven learning procedure merely under the assumptions that there exits causal relations among the 6 time series variables. (3) Importantly, the two Phillips curve equations confirm largely the theoretical formulation as given in (4.8) and (4.7), albeit some variables are statistically not significant: the price inflation and the wage inflation are driven by the common cost pressure variable π_t^m at different lags, both direct cost pressure dw_t and dp_t have no significant influence on the price inflation and the wage inflation respectively. A labour productivity increase dz_t will impact positively on the wage inflation with one lag, but has no impact on the price inflation. The market specific demand pressure e_t for wage inflation and u_{t-3} for price inflation have significant influence on dw_t and dp_t respectively.

An implication of the two Phillips curves is that the real wage dynamics can be conducted as difference of the two Phillips curves (See Flaschel and Krolzig (2003) for more detailed discussions on the stability of the real wage dynamics.).

$$dw_t - dp_t = 0.72\pi^m_{t-1} - 0.76\pi^m_t + 0.48e_t - 0.52u_{t-3} + 0.21dz_{t-1} + 7.25 + \epsilon_{wt} - \epsilon_{pt} \quad (4.15)$$

For the first two terms in the right hand side of the equation above it holds roughly $0.72\pi^m_{t-1} - 0.76\pi^m_t \approx -0.76(\pi^m_t - 0.96\pi^m_{t-1}) = -0.76\epsilon_{\pi t}$ the inflation expectation term will drop out of the real wage equation. We have a more concise real wage equation:

$$dw_t - dp_t = 0.47e_{t-1} - 0.52u_{t-3} + 0.21dz_{t-1} + 7.25 + \epsilon_t \quad (4.16)$$

The real wage equation above says that an increase in labor productivity in the previous period dz_{t-1} causes an increase in the current real wage. Beside the growth of labour productivity, a higher rate of labour utilization in the previous period e_{t-1} will also lead to an increase of the current real wage, but a higher rate of capacity utilization before three periods u_{t-3} will damp the growth of the current real wage. It is of interest to compare our derived formulation of the real wage dynamics with a traditional formulation of the real wage dynamics as given by equation (10) in Blanchard and Katz (1999). In their specification, the real wage growth depends on the growth of labour productivity, the rate of labour utilization and an error correction term of the difference between the lagged real wage and the lagged labour productivity. This real wage equation has been estimated for many OECD countries. For most European countries the error correction term appears significantly with a correct sign in the equation. For US data the error correction term is insignificant with a wrong sign. With our data set for Australia this error correction term is also insignificant. Without the error correction term, the real wage equation given in Blanchard and Katz (1999) is very similar to our formulation of the real wage equation. We have an additional rate of capacity utilization u_{t-3} as another influence variable on the real wage. In this sense, the our structural equations for the wage-price spiral derived by the data-driven causal inference is consistent with the findings in the literature.

In this analysis we come to the conclusion that both the wage inflation and the price inflation are driven by the inflation expectations. The temporal dependence between wage inflation and the price inflation, i.e. the mutual Granger causality is mainly the effect of the common causes: the inflation expectation, the capacity utilization, the labour utilization and the labour productivity growth. The estimated TSCM is very uninformative in providing an explanation how the inflation expectation is formed. It gives merely a AR1 process as a statistical description of the inflation expectation process. This is however not surprising, our model framework is designed to explain the wage price spiral but not the formation of inflation expectation, which will definitely need a more general theoretical framework including elements such as monetary policy and consumer behaviour. We will leave this for further research. Our empirical results show that although both the wage inflation and the price inflation are influenced by the inflation expectation, the real wage dynamic is not influenced by the inflation expectation. This a moderate support of the hypothesis of the classical dichotomy between real and nominal variables.

To assess the robustness of the inferred causal orders among the 6 variables, a bootstrap exercise was implemented. We use the estimated linear causal model (4.9) to (4.14) to generate bootstrap samples by bootstrapping the residuals. Based on the bootstrap residuals and the estimated TSCM we can generate bootstrap samples. Then we run the learning algorithm to obtain bootstrap partial DAGs. The frequencies of the inferred arrows in the bootstrap DAGs are reported in Table 3.

Arrow	Frequency	Arrow	Frequency
$u_{t-3} \to dp_t$	869	$dp_{t-3} \to dp_t$	5
$\pi^m_t \to dp_t$	1000	$u_{t-2} \to dp_t$	5
$\pi^m_{t-1} \to dw_t$	996	$\pi^m_{t-3} \to dw_t$	5
$dz_{t-1} \to dw_t$	965	$e_{t-4} \to dw_t$	21
$e_t \to dw_t$	623	$u_t \to dw_t$	5
$\pi^m_{t-1} \to \pi^m_t$	100	$dw_{t-2} \to dw_t$	5
$e_{t-3} \to e_t$	1000	$e_{t-2} \to dw_t$	5
$u_{t-3} \to e_t$	997	$u_{t-2} \to dw_t$	11
$e_{t-1} \to e_t$	1000	$dz_{t-2} \to dw_t$	5
$e_{t-2} \to u_t$	1000	$e_{t-2} \to dw_t$	8
$u_{t-1} \to u_t$	1000	$u_{t-2} \to dw_t$	13
$e_t \to u_t$	1000	$u_t \to dw_t$	5
$dz_t \to u_t$	935	$dw_{t-1} \to \pi^m_t$	5
$dz_{t-1} \to dz_t$	995	$e_{t-1} \to \pi^m_t$	5
		$dw_t \to \pi^m_t$	6
DAG	512	$e_t \to \pi^m_t$	7
		$u_{t-2} \to e_t$	9
		$dp_{t-1} \to u_t$	13
		$u_{t-1} \to dz_t$	5
		$\pi^m_t \to dz_t$	8

Notes: This table reports the frequency of identified arrows in the partial DAGs of the TSCMs estimated using 1000 bootstrap samples. We report the frequencies of all single arrows which are large than 4. Here we make no difference whether the arrows presenting causal direction or not. The frequencies of the true arrows are decisively larger than the frequencies of the wrongly estimated arrows.

Table 3: Frequency of Identified Causal Relations

The bootstrap results show clearly that all inferred causal relations (arrows) are very stable. The true arrows are all inferred with very high frequencies, while the wrongly identified arrows are with very low frequencies. Among the 14 arrows of the TSCM, 6 arrows are identified with a frequency of 100% 7 arrows are identified with frequencies close to one. Only one arrow $e_t \to dw_t$ has a frequency of 62.8%. This is responsible for the relatively low frequency of 51.2% in correctly identifying the complete partial DAG. All wrongly identified arrows with frequencies under 0.5% are not reported.

5 Concluding Remarks

In this paper we develop a method to uncover the causal relations in stationary multivariate time series. Grounded on the theory of inferred causation we embedded multivariate time series into a directed acyclic graphical model. Under the assumptions of the temporal causal constraint, the time-invariant causal structures constraint and the time-limited causal influence constraint, we define a time series causal model - TSCM. The DAG of a TSCM can be represented in a partial DAG. Combining a modified PC algorithm and a greedy search algorithm we are able to uncover the data-generating partial DAG effectively.

Our method extends the literature on causal analysis of VAR residuals to causal analysis of time series themselves and give the dynamics of multivariate time series a possible causal explanation. This method provides a new instrument to investigate the causal dependence among time series variables. Complex directionality of dependence among economic variables, such as uni-directional dependence, mutual dependence, temporal dependence and contemporaneous dependence can be represented in TSCMs. Further, a TSCM provides a causal explanation for the Granger causality among the time series in the TSCM.

We apply this method to analyze the wage-price spiral of the Australian economy for the period from 1978:3 to 2009:3. The mutual bivariate Granger causality between the price inflation and the wage inflation is the effect driven by common cause processes: the inflation expectation, the capacity utilization, the labour utilization and the productivity growth. Output of the learning algorithm includes the structural equations for the price inflation and the wage inflation, which gives not only the dependence among the variables but also directions of the dependence. It turns out that the two structural equations derived by the data-driven method of inferred causation confirm largely the structure equations postulated by theoretical consideration. The real wage dynamics implied by the two Philips curve structural equations gives a similar explanation to the real wage growth as it was given in Blanchard and Katz (1999). A bootstrap assessment of the robustness of the inferred causal order shows that our result is to a large extent reliable.

6 Appendix

Lemma 6.1 *Let A be the lower triangular adjacent matrix of a DAG. $A(i,j) = 1$ means there is an arrow from vertex j to vertex i and $A(i,j) = 0$ means there is no arrow between vertex j and vertex i. Let $A^{-1}(i,j)$ denote the (i,j) element of A^{-1}. $A^{-1}(i,j) \neq 0$ if and only if there exists a directed path from vertex j to vertex i.*

Proof: Since A is a triangular matrix it can be solved recursively. For a 4×4 matrix we obtain

$$
\begin{pmatrix}
1 & 0 & 0 & 0 \\
A_{21} & 1 & 0 & 0 \\
A_{31} & A_{32} & 1 & 0 \\
A_{41} & A_{42} & A_{43} & 1
\end{pmatrix}^{-1}
$$

$$
= \begin{pmatrix}
1 & 0 & 0 & 0 \\
-A_{21} & 1 & 0 & 0 \\
-A_{31} + A_{32}A_{21} & -A_{32} & 1 & 0 \\
-A_{41} + A_{42}A_{21} + A_{43}A_{31} - A_{43}A_{32}A_{21} & -A_{42} + A_{43}A_{32} & -A_{43} & 1
\end{pmatrix}.
$$

It is easy to see that the (i, j) element in A^{-1} consists of summands each of whom represents a directed path from j to i. For example $A^{-1}(3, 1) = -A_{31} + A_{32}A_{21}$. $A_{3,1}$ represents an arrow $1 \to 3$ and $A_{32}A_{21}$ represents a directed path consisting of the arrows $1 \to 2 \to 3$. These two directed paths are the only paths from vertex 1 to vertex 3. For $A^{-1}(4, 1) = -A_{41} + A_{42}A_{21} + A_{43}A_{31} - A_{43}A_{32}A_{21}$. The summands in the sum represent the directed paths from vertex 1 to vertex 4: $1 \to 4$, $1 \to 2 \to 4$, $1 \to 3 \to 4$ and $1 \to 2 \to 3 \to 4$ respectively and they include all possible paths from vertex 1 to vertex 4. It is easy to verify that for an $N \times N$ matrix the same rule holds. Assuming that all element in A are unconstrained varying parameters, the sum will be zero, if and only if each summand in the sum is zero. This implies that if there is a path from j to i then the (i, j) element of A^{-1} will be nonzero. $\square$

Proof of Proposition 2.8

Given the correspondence between VAR (2.5) and TSCM (2.4), we have the relation

$$
\Pi_s(i, k) = \sum_{j=1}^{N} A_0^{(-1)}(i, j)A_s(j, k) = \sum_{j=1}^{j} A_0^{(-1)}(i, j)A_s(j, k). \tag{6.17}
$$

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 the (j, k) element of the TSCM coefficient matrices A_0^{-1} and A_s, respectively. The second equality in equation (6.17) is because that the inverse of A_0 is a lower triangular matrix. In the VAR framework $X_{k,t}$ is a Granger cause of $X_{i,t}$ if and only if $\Pi_s(i, k) \neq 0$ for some $s > 0$. If there is a path from $X_{k,t-s}$ to $X_{i,t}$ via $X_{j,t}$ we have $A_0^{(-1)}(i, j)A_s(j, k) \neq 0$. If there is a direct path from $X_{k,t-s}$ to $X_{i,t}$ we have $A_0^{(-1)}(i, i)A_s(i, k) \neq 0$. Both cases imply $\Pi_s(i, k) \neq 0$, i.e. $X_{k,t}$ is a Granger cause of $X_{i,t}$.

If $X_{k,t}$ is a Granger cause of $X_{i,t}$, we have $\Pi_s(i, k) = \sum_{j=1}^{n} A_0^{(-1)}(i, j)A_s(j, k) \neq 0$, for some $s > 0$. This implies that there is either one pair $A_0^{(-1)}(i, j)$ and $A_s(j, k)$ are nonzero, or $A_s(i, k)$ is nonzero. This corresponds to that there is either a directed path from $X_{k,t-s}$ to $X_{i,t}$ via $X_{j,t}$ or a direct arrow from $X_{k,t-s}$ to $X_{i,t}$. $\square$

References

Blanchard, O. J. and Katz, L. (1999). Reconciling theory and evidence. *American Economic Review*, 89(2):69–74.

Cartwright, N. (2001). What is wrong with Bayes nets? *MONIST*, 84:242–264.

Chen, P. and Flaschel, P. (2006). Measuring the interaction of wage and price Phillips curves for the U.S. economy. *Studies in Nonlinear Dynamics and Econometrics*, 10, No. 4:Article 2.

Chen, P. and Hsiao, C. (2007). Learning causal relations in multivariate time sereis data. *Economics: The Open-Access, Open-Accessment E-Journal*, 1, 2007-11.

Demiralp, S. and Hoover, K. (2004). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65:745–767.

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

Flaschel, P. and Krolzig, H. (2003). Wage and price Phillips curves. An empirical analysis of destabilizing wage-price spirals. *Center of Empirical Macroeconomics, Bielefeld University*.

Freedman, D. and Humphreys, P. (1998). Are there algorithms that discover causal structure? *http://www.stat.berkeley.edu/ census/514.pdf*.

Glymour, S. and Spirtes, G. (1988). Latent variables, causal model and overidentfying constraints. *Journal of Econometrics*, 39:175–198.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, 1st edition.

Heckerman (1995). A tutorial on learning with bayesian networks. *Microsoft Research*, MSR-TR-95-06:http://research.microsoft.com/research/pubs.

Hoover, K. (2005). Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory*, 21:69–77.

— (2008). Causality in economics and econometrics. *The New Palgrave Dictionary of Economics Online, Second Edition. Steven N. Durlauf and Lawrence E. Blume Eds. Palgrave Macmillan*.

Pearl, J. (2000). *Causality*. Cambridge University Press, 1st edition.

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 Rea- soning: Procedings of the 2nd International Conference, San Mateo, CA: Morgan Kaufmann*, pages 441–452.

Robins, J., Scheines, R., Sprites, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90:941–515.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. Springer-Verlag, New York / Berlin / London / Heidelberg / Paris, 2nd edition.

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.