Expert Systems with Applications 36 (2009) 12441–12445

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Automatic linear causal relationship identification for financial factor modeling

Zitian Wang *, Shaohua Tan

Department of Intelligence Science, Center for Information Science, Room 2314, Science Building 2, Peking University, Beijing 100871, China

ARTICLE INFO

Keywords:
Causality
Causal Discovery
SEM
Financial factors

ABSTRACT

Given a comprehensive set of financial factors, we use linear non-Gaussian SEM to automatically identify the causal relationships buried in the factor set. The causal structure is allowed to have cyclic edges, explicitly accommodating 'mutual causality' which is well acknowledged but rarely modeled in standard economic theory. The method takes advantage of both artificial intelligence and economic related techniques, and identifies one stable model from several distribution-equivalent equilibrium models for each dataset. Empirical studies on 15 financial factors reveal some interesting findings, especially for the risk-return relationship modeling and capital structure determinants discovery.

1. Introduction

The financial market is one of the most complicated systems and is now undergoing constant innovations. There are a lot of factors that measure different aspects of the financial status of a firm, such as size, profit, debt level, firm growth prospect, and operation efficiency. In conventional financial research, models are usually small-scaled with only a small set of financial features studied. For example, in identifying the determinants of asset return, Naranjo, Nimalendran, and Ryngaert (1998) study the relationship only from the dividend yields aspect, while Jacobs and Wang (2004) considers only consumption risk. These studies, standing on different points of view, rarely reach a consensus and the key relationships which may show up from an extended viewpoint tend to be disregarded. Recently, new researches are emerging to work on a comprehensive set of factors and build up models from a global perspective (De Mol, Giannone, & Reichlin, 2006; Wang, Wang, & Tan, 2008). In this paper, we also take a global view in studying the causal relationships of financial factors, with the aim to gain deeper understating of essence of the market, and discover more rules that link financial causes and effects.

For the relationship study of a set of financial factors, one of the powerful tools is the causal discovery technique, which is attracting increasing interest from different communities. In the Artificial Intelligence literature, Bayesian networks (BNs) are developed in learning the directed acyclic graphical structure representing the variable causal relationships (Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993). The exact bayesian network structure can be learned for relatively small dimensions (Koivisto & Sood, 2004; Koivisto, 2006), while for large dimensions there are efficient and effective structure learning algorithms (Tsamardinos, Brown, & Aliferis, 2006). Another main stream of research is from the econometric modeling of causality (Granger, 1969; Heckman & Vytlacil, 2005; Hoover, 2001, 2005, 2008; Simon, 1953; Sims, 1972, 1980). The structural equations modeling (SEM) approach originated from the Cowles Commission provides methods to help understanding and measuring economic causal effects. The structural vector autoregression, closely related with Granger's model, are developed for dynamic time-series casuality analysis.

In this paper, based on a comprehensive set of financial factors, we use the newly developed linear non-Gaussian SEM (Lacerda, Spirtes, Ramsey, & Hoyer, 2008) to automatically identify the causal structure from data. The learned causal relationships are encoded in a set of structural equations, and represented in the corresponding graph. The graph is allowed to contain cyclic edges, explicitly accommodating mutual causality which is well acknowledged but rarely modeled in economic theory.

The linear non-Gaussian SEM method that we use is different from, but meanwhile takes advantage of techniques of both artificial intelligence and the economic approaches mentioned above. In our setting, the structure for the financial factors is learned from data. The causal discovery process reveals the influential relationships buried in the composite factor set. This learning mechanism is in contrast to traditional SEM where the factor relationships are totally specified by the economic experts based on their hypotheses or instincts. Moreover, the existence of cyclic edges in the structure extends the traditional Bayesian network acyclic graphical approach. With cycles, the graph encodes the set of variables in equilibrium. Furthermore, the basic SEM forms are retained in the factor relationship representation, which is consistent with the main-stream economic thinking.

From the economic point of view, through comprehensive factor set analysis we can identify the influential factors of the *Rate*

* Corresponding author. Tel.: +86 10 62755745.
E-mail addresses: wzt@cis.pku.edu.cn, princestein2008@hotmail.com (Z. Wang).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.04.056

of return, which are the'risk factors' in the economic literature. Finding the risk factors and building up the risk-return relationship has been of central interest of financial studies (Bodie, Marcus, & Kane, 2002; Sharpe, 1964). Apart from that, we can also analyze the interrelationships of the risk factors, which is seldom discussed in conventional economic modeling. Closely related with the *capital structure* theory there are also two central factors: *Return on Assets* and *Gross Margin*. The graph helps specifying the determinants of capital structure, which provides a refinement of thinking in how the firm characteristics affect the capital structure choice. What is more, the analysis can be carried on for consecutive years by monitoring the similarities and dynamics of the structures, which would help finding the patterns in the market that last for a long period and identifying the specific characteristics of each year.

The rest of the paper is organized as follows: in Section 2 we introduce the financial factor causal discovery problem and the linear non-Gaussian SEM method. In Section 3, based on cross-sectional data of a set of 15 financial factors, we illustrate the empirical results among which there are interesting findings discovered. Conclusions and discussions are in Section 4.

2. Causal discovery from financial factors

In general, the aim of the study can be summarized as follows: given the financial factors $\mathbf{F} = \{F_1, F_2, \ldots, F_d\}$ in which d is the dimension of the factor set, we try to find a set of simultaneous equations for the factors in which the causal relationships are encoded.

Suppose data samples $\mathbf{x}$ are collected for $\mathbf{F}$, where $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}^T$ and $x_i (1 \leqslant i \leqslant d)$ is a $1 \times n$ vector, n is the number of samples.

First we assume that

Assumption 1. Each factor is affected by a limited number of other factors, with the relationship described by linear equations.

This generic linear assumption dominates the main stream of economic thinking. For example in Arbitrage Pricing Model (Ross, 1976) in finance, the expected *Rate of return* of an asset is modeled as a linear function of various macroeconomic or company specific factors. We follow the same line of thinking in this paper, assuming the linear relationship among factors.[1]

The financial system is a kind of dynamical systems undergoing constant changes and adjustments. The listed firms are periodically reporting financial statements to provide reliable and comparable information of the firm's financial position and performance. The factors we study are mostly collected from the financial reports, and it is assumed that these factors are from the equilibrium state of the market, and their values will not change in the absence of external influences:

Assumption 2. The factors are from the market equilibrium.

It should be noted that "Not all economic equilibria are stable". For an equilibrium to be stable, the economic forces would return to their original equilibrium state after a minor disturbance. In the situations when equilibrium state can not be reached after a small deviation, they are the unstable equilibria.

Given Assumptions 1 and 2, the observed equilibria can be expressed in the form of structural equations model

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{1}$$

in which matrix B encodes the structure of the equilibria, and $\mathbf{e}$ is the vector of errors.

[1] It is straightforward to extend the linear non-Gaussian SEM to the nonlinear version using kernel methods (Shawe-Taylor & Cristianini, 2004).

Next we propose a assumption on the error terms $\mathbf{e}$.

Assumption 3. For $\mathbf{e} = \{e_1, \ldots, e_d\}^T$, the error terms $e_i (1 \leqslant i \leqslant d)$ are statistically independent and at most one error term is Gaussian.

The statistical independence assumption implies that we have collected sufficient variables and there are not unobserved common causes of any two distinct factors in the factor set. This explains why we need to take a comprehensive view and collect various factors that cover different aspects of the financial status.

Unlike the traditional Gaussian assumption of the error terms, we assume that the errors are mostly non-Gaussian. This would help identifying the direction of the causal relationship. To explain it, we give a simple bivariate example as shown in Fig. 1.

For model 1 in Fig. 1, the underlying data generating process is

$$\begin{aligned} y_t &= \alpha x_t + \epsilon_{1t} \\ x_t &= \epsilon_{2t} \end{aligned} \tag{2}$$

and for model 2 the data generating process is

$$\begin{aligned} y_t &= \omega_{1t} \\ x_t &= \beta y_t + \omega_{2t} \end{aligned} \tag{3}$$

If the error terms ϵ_{it} and ω_{it} are Gaussian, then from observational data alone, model 1 and model 2 can not be distinguished, i.e., they are observational equivalent. But if at least one term of ϵ_{it} and ω_{it} is non-Gaussian, then it is easy to verify that the two models are different (Hoyer, Shimizu, Kerminen, & Palviainen, 2008).

Given Assumptions 1, 2, and 3, we use the recently developed linear non-Gaussian SEM method (Lacerda et al., 2008) to discover the causal structure of financial factors.

The causal discovery process is briefly introduced as follows. From Eq. (1), there is $\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$. Denoting $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, we get the reduced form

$$\mathbf{x} = \mathbf{A}\mathbf{e} \tag{4}$$

Since we only know $\mathbf{x}$, there are infinite pairs of invertible $\mathbf{A}$ and corresponding $\mathbf{e}$ which satisfy Eq. (4). Since it is assumed that the error terms e_i are statistically independent in Assumption 3, practically we try to find the pair of $\mathbf{A}$ and $\mathbf{e}$ for which the $e_i s$ are maximally independent. The Independent Component Analysis (ICA) (Comon, 1994; Hyvarinen, Karhunen, & Oja, 2001) is used to solve the problem. There are well developed ICA algorithms and for clarity, details of ICA implementations are left out here for reference in Hyvarinen (1999), Zhang and Chan (2008).

The ICA algorithm returns the estimate of $\mathbf{W} = \mathbf{A}^{-1}$, namely, $\mathbf{W}_{\mathrm{ICA}}$. But there is one deficiency of ICA: the returned $e_i s$ are in arbitrary ordering. So we need to reorder the rows of $\mathbf{W}_{\mathrm{ICA}}$ to find the correct correspondence between the error terms $\mathbf{e}_i$ and the factor x_i.

To find the 'correct' correspondence, there are two more restrictions imposed. First, the should be no 1's on the diagonal of esti-

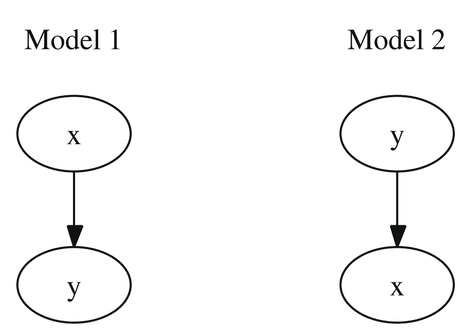

Fig. 1. A bivariate model.