XVIII

Volume XVIII, Number 1, May 2015

Journal of Applied Economics

Pei-Chun Lai
David A. Bessler

Price discovery between carbonated soft drink manufacturers and retailers: A disaggregate analysis with PC and LiNGAM algorithms

UCEMA

Edited by the Universidad del CEMA
Print ISSN 1514-0326
Online ISSN 1667-6726

Journal of Applied Economics. Vol XVIII, No. 1 (May 2015), 173-198

PRICE DISCOVERY BETWEEN CARBONATED SOFT DRINK MANUFACTURERS AND RETAILERS: A DISAGGREGATE ANALYSIS WITH PC AND LiNGAM ALGORITHMS

Pei-Chun Lai*
Capital University of Economics and Business

David A. Bessler
Texas A&M University

Submitted October 2013; accepted August 2014

This paper considers the use of two machine learning algorithms to identify the causal relationships among retail prices, manufacturer prices, and number of packages sold. The two algorithms are PC and Linear Non-Gaussian Acyclic Models (LiNGAM). The dataset studied comprises scanner data collected from the retail sales of carbonated soft drinks in the Chicago area. The PC algorithm is not able to assign direction among retail price, manufacturer price and quantity sold, whereas the LiNGAM algorithm is able to decide in every case, i.e., retail price leads manufacturer price and quantity sold.

JEL classification codes: C1, L1
Key words: causal relationships, PC algorithm, Linear Non-Gaussian Acyclic Models (LiNGAM)

I. Introduction

This paper addresses price discovery at the interface between two manufacturers and one retailer of carbonated soft drink (CSD) products. We use two algorithms from the field of machine learning: the PC algorithm and a recently developed Linear Non-Gaussian Acyclic Model (LiNGAM). We apply both of them to infer

* Pei-Chun Lai (corresponding author): International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China; e-mail: pclaitamu@gmail.com. David A. Bessler: Department of Agricultural Economics, Texas A&M University, College Station, TX 77843, USA; e-mail: d-bessler@tamu.edu. We are indebted to Shohei Shimizu and Patrik O. Hoyer for helpful discussions. We also would like thank Conte Grand (the Editor), and two anonymous referees for many valuable comments. The authors would like to gratefully acknowledge the support of the Capital University of Economics and Business 2013 Research Start-up Funding.

the causal relationship, if any, among retail price, manufacturer price, and quantity (number of packs) sold based on scanner data collected from the sale of carbonated soft drinks in the Chicago area.

We begin by summarizing the literature on manufacturer-retailer channel interactions and the corresponding channel power. Gardner (1975), who offers a theoretical treatment of the farm-retail price spread for food products, shows how the details of underlying demand and supply information affect the market-generated retail-farm price spread. One drawback is that the Gardner treatment with observational data requires considerable knowledge of the underlying demand and cost (supply) structure. Haines (2007) shows that the manufacturer's brand share and brand price premium have significant negative impacts on the amount a retailer buys and sells on promotion. One possible reason is that if more powerful brands offer fewer promotions and with smaller discounts, the percentage of brand volume the retailer buys or sells on promotion will be lower for the higher share brand. Moreover, the retailer share has a significant positive impact on the percentage bought by the retailer on promotion. Haines, however, only examines the influence of manufacturer or retailer power, i.e., manufacturer's brand share, price premium or retailer share, on retailer response to trade discounts, and does not provide a method to measure the division of channel profits between manufacturer and retailer, where a higher share of channel profit is associated with greater channel power. Villas-Boas (2007), who investigates vertical relationships between manufacturers and retailers of yogurt when wholesale price data is unavailable, computes price-cost margins for retailers and manufacturers that are implied by alternative vertical contracting models and compares the margins with the estimated price-cost margins by using components of marginal costs to assess the fit of different vertical models. The result is consistent with the high bargaining power of retailers that are able to force wholesale prices down to marginal cost and thus the manufacturers' implied price-cost margins are 0 for all products. Kadiyali et al. (2000) use a conduct parameter approach to address vertical relationships in the juice and tuna marketing channels through the notion of pricing power. Their empirical results demonstrate that the retailer has greater pricing power in the channel. The above approaches make heavy use of assumptions about the exact specification of structural equations. We note that these estimations are sensitive to misspecification in any equation. All of these assumed structural models need to be realistic, because an incorrect choice of functional form leads to biased and unreliable results (Huang, Rojas, and Bass 2008).

Our paper explores two machine learning algorithms that attempt to uncover causal relationships between or among variables without assuming causal relationship *a priori*. The algorithms are constrained to search over acyclic graph structures and the resulting output is termed a Directed Acyclic Graph (DAG). DAG is widely used to represent causal relationships among non-temporal variables (Pearl 2009). Several algorithms have been used to generate DAG for the purpose of correctly describing the causal association among variables. The PC algorithm, an early contribution to the field, makes the assumption of Gaussian data and applies a conditional independence relation. The assumption of Gaussian data negates the need for information of higher-order moment structures (Shimizu, Hoyer et al. 2006). The PC algorithm, however, often leads to a set of indistinguishable causal patterns that are equivalent in their conditional probability structure (or probability structure). For example, when x, y, and z are normally distributed, the data cannot distinguish between the two graphs: $x \leftarrow y \rightarrow z$ and $x \rightarrow y \rightarrow z$. In other words, both graphs are compatible with the same probability distribution, and therefore are indistinguishable and observationally equivalent (Pearl 2009). A key character of LiNGAM is the assumption of non-Gaussianity of the variables (or disturbances) (Shimizu, Hoyer et al. 2006). If the data are non-Gaussian, we can apply higher-order moment structures to identify causal patterns, some of which may not be distinguished by the PC algorithms. The further the data are from normality, the more accurate the ultimate causal patterns identified by LiNGAM (Shimizu and Kano 2008). In this paper, we elucidate the differences and similarities between the PC and LiNGAM algorithms. We reject normality for all of our data, and thus LiNGAM becomes a suitable candidate for identifying any underlying causal structure among our variables.

II. PC algorithm

A. Conditional independence statements and graphs

The PC algorithm, one of the earliest and widely-used machine learning algorithms, is based on the concept of conditional independence. For explanatory purposes, we describe the concept of a dependency model. Let X, Y, and Z denote three disjoint subsets of variables. "X is independent of Y given Z" can be denoted by the independency statement, $I(X, Y|Z)$. Suppose M is a dependency model, which is a rule that determines whether $I(X, Y|Z)$ is true. If there is a direct correspondence

between the variables of M and the set of vertices in V of an undirected graph $G=(V,E)$[1], then the topology of G reflects some properties of M. A subset Z of nodes in a graph G that intercepts all paths between the nodes of X and those of Y can be written as $\langle X \mid Z \mid Y \rangle_G$. When two sets of nodes X and Y are connected through a set Z, conditioning on Z can be understood as blocking these interactions (Pearl 1988; Kwon and Bessler 2011). This leads to the following definition.

Definition 1 *(Pearl 1988) An undirected graph G is a dependency map (or dependence map, or D-map) of a dependency model M if there is a one-to-one correspondence between the variables of M and the nodes V of G, such that for all disjoint subsets X, Y, and Z of elements*

$$I(X,Y \mid Z)_M \Rightarrow \langle X \mid Z \mid Y \rangle_G .$$

Similarly, G is an independency map (or independence map, or I-map) of M if

$$I(X,Y \mid Z)_M \Leftarrow \langle X \mid Z \mid Y \rangle_G .$$

G is said to be a perfect map (P-map) of M if it is both a D-map and an I-map. Therefore

$$I(X,Y \mid Z)_M \Leftrightarrow \langle X \mid Z \mid Y \rangle_G .$$

Any probability distribution P is a dependency model, because for any triplet *(X, Z, Y)* the validity of $I(X, Y|Z)$ can be tested using the following equation $P(x| y,z) = P(x \mid z)$whenever $P(y,z) > 0$, where *x, y,* and *z* represent the assigned values of the variables *X, Y*, and *Z,* respectively (Pearl 1988). Given a probability distribution P that satisfies the Causal Markov Condition and Stability condition, a DAG (or Bayesian Networks) G is a perfect map of P for the continuous normal distribution and for the discrete multinomial distribution (Pearl 1988; Kwon and Bessler 2011).

[1] *G=(V,E)* is a graph consisting of nodes *V* in one-to-one correspondence with the variables, and edges (lines) *E* that connect the nodes. By undirected we mean there is no arrowhead at the end of the line indicating causal influence (*X* causes *Y*), but merely a line, $X-Y$, where *X* and *Y* are related, but the direction of causal flow is unknown. A directed graph has vertices whose edges are connected with arrows, e.g., $X \rightarrow Y$. A directed acyclic graph is a directed graph that contains no directed cyclic paths. A path represents a sequence of consecutive edges in the graph and blocking can be interpreted as stopping the flow of information (or dependency) between the variables that are connected by the paths (Pearl 2009).

B. Causal Markov Condition and *d*-separation

The common distribution assumption used in the PC algorithm is Gaussian distribution for continuous variables.[2] In a more generalized way, a parameterized DAG for a set of variables is a pair (G, Θ_G) where G is a DAG and Θ_G is the set of free parameters mapping the graph onto a probability distribution. We note that can be the coefficients as well as the means and variances of the error terms of the structural equations.[3] In order for the parameters to specify a probability distribution, some restrictions must be imposed on the parameters, e.g., the standard deviations cannot be negative. Any parameter value that falls within the restricted range is termed a "legal" parameter value. $P(\langle G, \Theta_G \rangle)$ denotes the set of all distributions corresponding to legal parameter values. $I(\langle G, \Theta_G \rangle)$ denotes the set of conditional independence relations that holds in every distribution in (Spirtes 2005). A DAG G represents any joint distribution over the variables $X=\{X_1,...,X_n\}$ that can be factored according to the following rule

$$p\left(X_1 = x_1,\ldots, X_n = x_n\right) = \prod_{i=1}^{n} p\left(X_i = x_i \mid Pa_i^G = pa_i^G, \Theta_G\right), \tag{1}$$

where $\boldsymbol{Pa_i^G}$ is the set of parents (direct causes) of node x_i in G. The factorization of p according to G is equivalent to each variable X in the DAG being independent of all the variables that are neither parents nor descendants of X, and conditional on all of the parents of X in G. Any probability distribution that satisfies the property in Equation (1) is said to satisfy the Causal Markov Condition for G (Spirtes 2005). For any $\boldsymbol{P}(\langle \boldsymbol{G}, \boldsymbol{\Theta_G} \rangle)$ satisfying the Causal Markov Condition for G, all of the conditional independence relations in $\boldsymbol{I}(\langle \boldsymbol{G}, \boldsymbol{\Theta_G} \rangle)$ hold. Pearl has proposed the concept of *d*-separation to determine which conditional independence relations are entailed by satisfying the Causal Markov Condition. A path is said to be *d*-separated by *Sepset*(X) (separating set) if and only if (iff) (1) a path contains a causal chain $X_1 \rightarrow X_2 \rightarrow X_3$ or causal fork $X_1 \leftarrow X_2 \rightarrow X_3$ such that X_2 should be in the *Sepset* (X_1,X_3) because X_1 and X_3, which are unconditionally dependent, become independent once conditioned on X_2, or (2) a path contains an inverted fork (or

[2] A reviewer has properly pointed out that one could actually use more general tests of independence and conditional independence to implement the PC algorithm. In this paper our use of the PC algorithm follows the current implementation as given in TETRAD (http://www.phil.cmu.edu/projects/tetrad/) under the assumption of Gaussianity.

[3] The structural equation $X \rightarrow Y$ of can be written as $Y = \theta_{xy} X + \varepsilon_y$.

unshielded collider, or *v*-structure) $X_1 \rightarrow X_2 \leftarrow X_3$ such that $X_2 \notin Sepset\ (X_1, X_3)$ and any of X_2's descendants[4] are not in $Sepset\ (X_1, X_3)$ because X_1 and X_3, which are unconditionally independent, become dependent once conditional on X_2or its descendants (Spirtes et al. 2001; Pearl 2009). If $I\ (X_1, X_3 |\ X_2)_p$ is satisfied whenever X_1 is *d*-separated from X_3 conditional on X_2 in G, then P satisfies the Causal Markov Condition (Spirtes 2010). Following Drewek (2010), the PC discovery algorithm can be summarized as shown in Algorithm 1.[5]

Table 1. Algorithm 1: PC discovery algorithm

1. Form a complete undirected graph G on the vertex set $X = \{X_1, \ldots, X_n\}$.

2. Set k=0 and repeat the following two steps

   i. Test for all ordered pairs of adjacent vertices (X_i, X_j) in G with $|Adjacencies(X_i) \setminus X_j| \geq k$ if a subset $S \subset Adjacencies(X_i) \setminus X_j$ exists that fulfills

   $$|S| = k \text{ and } (X_i \perp X_j)|S$$

   If so, remove edge (X_i, X_j) from G and save S as $Sepset(X_i, X_j)$

   ii. Set k=k + 1 and go back to 2i.

3. For each ordered triple of vertices (X_i, X_j, X_k) such that pairs (X_i, X_j) and (X_j, X_k) are adjacent, but (X_j, X_k) is not adjacent, check if $X_j \in Sepset(X_i, X_k)$. If so, orient the edges as $X_i \rightarrow X_j \leftarrow X_k$.

4. Orient as many of the remaining undirected edges, such that neither a new *v*-structure nor a directed cycle is created.

However, even given the Causal Markov and Faithfulness Assumptions and the assumption of sufficiency,[6] we note that the true causal model is underdetermined because of the hierarchy of equivalence relations. For example, causal chain and

[4] The edge $X \rightarrow Y$ represents X is a parent of Y. Y is a descendant of X if there is a directed path from X to Y (Spirtes 2010).

[5] $|Adjacencies(X_i) \setminus X_j|$ Calculates the cardinality of the neighborhood of the X_i's edge without X_j.

[6] Suppose $P(\langle G, \Theta_G \rangle)$ is represented by a DAG G, then P is stable (or faithful) to G iff $I(\langle G, \Theta_G \rangle)$ is entailed (by d-separation) by G for any set of the free parameters. In other words, the stability condition states that no $I(\langle G, \Theta_G \rangle)$ can be destroyed as we vary the parameters from Θ_G to Θ'_G (Spirtes 2005; Pearl 2009). A set of variables X is causally sufficient iff the error terms are mutually independent. The system does not omit any variables that are direct causes of any pair of variables in X (Spirtes et al. 2004).

causal fork represent exactly the same set of probability distributions so these two DAGs are distributionally equivalent. In addition, $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$ are conditional independence equivalent iff both graphs entail the same set of conditional independence relations, $\boldsymbol{I}(\langle \boldsymbol{G_1}, \boldsymbol{\Theta}_{\boldsymbol{G_1}} \rangle) = \boldsymbol{I}(\langle \boldsymbol{G_2}, \boldsymbol{\Theta}_{\boldsymbol{G_2}} \rangle)$ (i.e., they have the same set of *d*-separation) (Spirtes et al. 2004; Spirtes 2005). In the case of multivariate normal distributions, under the condition of sufficiency, conditional independence equivalence does entail distributional equivalence (Spirtes 2005). In most of the PC results, more than one causal graph is conditional independence equivalent and is compatible with a given probability distribution (distributional equivalent), and thus, without further background knowledge, no reliable statistical inference from the data can distinguish between them (e.g., causal chain and causal fork). Therefore, the resultant graph is not unique (Moneta et al. 2013). Such equivalences are characterized by undirected edges in the graph (Drewek 2010).

III. Linear Non-Gaussian Acyclic Models (LiNGAM)

In general, the PC algorithm searches the causal pattern based on conditional independence, whereas the LiNGAM algorithm discovers the causal directionality based on functional composition (Pearl 2009). LiNGAM identification relies on independent component analysis (ICA).[7] For Gaussian variables, ICA cannot find the correct mixing matrix because many different mixing matrices yield the exact same Gaussian joint density (Hyvärinen et al. 2001). ICA is only feasible on non-Gaussian data. We can use higher-order statistics of the variables to obtain stronger identification results if the data are non-Gaussian.[8] The details are as follows.

A. Independent component analysis (ICA)

The Central Limit Theorem (CLT) states that any mixture of independent source signals usually has a distribution that is closer to a normal distribution than any

[7] We apply ICA-LiNGAM in this paper. Shimizu et al. proposed the second estimation algorithm for LiNGAM which is known as DirectLiNGAM (Shimizu et al. 2011). DirectLiNGAM is an alternative estimation method that does not make use of ICA.

[8] In the PC algorithm, the variables are assumed Gaussian. Under the assumption of Gaussianity, we cannot use higher-order moment structure to identify the causal direction between two variables because the data are not skewed (Dodge and Rousson 2001).

of the constituted original variables (Stone 2004). Assuming that we observe the mixtures, $X=(X_1,...,X_n)$ of the independent signals $S=(S_1,...,S_n)$ [9], we have

$$X = As, \tag{2}$$

where s are mutually independent components. According to the CLT, any of the s is less Gaussian than the mixture variables X. We can rewrite the independent components as the linear combination of the mixture variables inversely. The objective of ICA is to find the "demixing matrix" W where W maximizes the sum of the non-Gaussianity of the mutually statistically independent components of $\tilde{S}$ where $\tilde{S} = \tilde{W}X$ and $\tilde{W} = \boldsymbol{A}^{-1}$(Hyvärinen et al. 2001; Shimizu, Hyvärinen et al. 2006).

B. Linear Non-Gaussian Acyclic Models (LiNGAM)

Shimizu, Hoyer et al. (2006) developed LiNGAM to implement a causal search on non-Gaussian distributed variables based on the assumption of independently distributed non-Gaussian disturbances.[10] Assume that causal relationships exist among the vector $X=(X_1, X_2,...,X_n)$ and can be represented by the structural equation model

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i \tag{3}$$

where $k(i)$ denotes a causal order of x_i and X_j is a direct cause of X_i. The disturbances e_i are mutually independent and non-Gaussian distributed with non-zero variances. If each variable x_i has a zero-mean, we are left with the following system of equations:

[9] The original assumption of the ICA model is that the number of observed variables must be greater than or equal to the number of independent signals.

[10] More details about LiNGAM appear in Shimizu, Hyvärinen et al. (2005) and Shimizu, Hoyer et al. (2006).

$$X = \mathrm{B}X + e, \tag{4}$$

where B is the coefficient matrix of the model. Solving for X in equation (4) gives

$$(I - \mathrm{B})X = e \quad \Rightarrow X = (I - \mathrm{B})^{-1} e = Ae, \tag{5}$$

that is, equation (5) and the above non-Gaussianity of disturbances form the classical linear ICA model (Hyvärinen et al. 2001; Shimizu, Hyvärinen et al. 2006).The error terms e in equation (5) can be viewed as sources or signals s. We can rewrite equation (5) as

$$e = (I - B)X = \tilde{W}X. \tag{6}$$

In general, the LiNGAM algorithm is processed by first conducting ICA estimation[11] to estimate the mixing matrix A, and then permuting and normalizing it appropriately before computing B.[12] Following Shimizu, Hoyer et al. (2006), the discovery algorithm can be briefly summarized as shown in Algorithm 2.

Table 2. Algorithm 2: ICA-LiNGAM discovery algorithm

1. Given an m x n data matrix X, where each column contains one sample vector x, subtract the mean from each row of X, apply an ICA algorithm to obtain a decomposition $X = AS$, and calculate $W = A^{-1}$.
2. Find a permutation of rows of W yielding a matrix $\widetilde{W}$ without any zeros on the main diagonal.
3. Divide each row of $\widetilde{W}$ by its corresponding diagonal element to yield a new matrix $\widetilde{W}'$ with all ones on the diagonal.
4. Compute an estimate $\widetilde{B} = I - \widetilde{W}$ of B.
5. Permute $\widetilde{B}$ until it is strictly lower triangular to gain the causal order.

[11] In most cases, ICA decomposition is implemented by using the FastICA algorithm.
[12] More details of permutation and normalization appear Shimizu and Kawahara (2010).

Note that some remaining estimated edges between variables may be weak and are probably zero in the generating model. The Wald test can be used to determine if some remaining connections should be pruned.[13]

C. Determining the direction of causality[14]

Suppose
Model 1: $y = \beta x + \varepsilon_y$,
Model 2: $x = \eta y + \varepsilon_x$,

where the explanatory variable is independent of the error in each model. Let x_k and $y_k (k = 1,...,N)$ be observations on x and y with mean zero. Define the moment structure as

$$m_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_k^i y_k^j$$

Because $E(x) = E(y)=0$, we do not consider the first-order moment of observed data.

The model-predicted second-order moment structure of Model 1 is

$$E\begin{bmatrix} m_{20} \\ m_{11} \\ m_{02} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \\ \beta^2 & 1 \end{bmatrix} \begin{bmatrix} E\left(x^2\right) \\ E\left(\varepsilon_Y^2\right) \end{bmatrix} = \sigma_2\left(\hat{\tau}_2\right), \tag{7}$$

where τ_2 is the number of parameters, in this case, $\hat{\tau}_2 = \left(\beta, E\left(x^2\right), E\left(\varepsilon_Y^2\right)\right)$. Note that the number of the distinct sample moments and the number of τ_2 are both 3. Also note that Model 1 and Model 2 have the same second-order moment structure. Under the above conditions, therefore, Model 1 and Model 2 are equivalent, which means that Model 1 cannot be identified from Model 2 if we only consider second-

[13] The detail of pruning edges appears in Shimizu, Hoyer et al. (2006).
[14] The detail of finding the causal direction between non-Gaussian x and y appears in Kano and Shimizu (2003), and Shimizu and Kano (2008). Our treatment in this section closely follows these original authors.

order moment structures. However, if the relevant variables and disturbance terms are non-normally distributed, we can apply the higher-order moments of Model 1 and Model 2 to detect the causal direction. For example, the third-order moment of Model 1 can be expressed as

$$E\begin{bmatrix} m_{30} \\ m_{21} \\ m_{12} \\ m_{03} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \\ \beta^2 & 0 \\ \beta^3 & 1 \end{bmatrix} \begin{bmatrix} E\left(x^3\right) \\ E\left(\varepsilon_Y^3\right) \end{bmatrix} = \sigma_3\left(\hat{\tau}_3\right), \tag{8}$$

where $\hat{\tau}_3 = \left(\beta, E\left(x^3\right), E\left(\varepsilon_Y^3\right)\right)$. The fourth-order moment structure can be defined in a similar way. Let

$$m = \left[m_2^T, m_3^T, m_4^T\right]^T \text{ and } \sigma(\tau) = \left[\sigma_2(\tau_2)^T, \sigma_3(\tau_3)^T, \sigma_4(\tau_4)^T\right]^T.$$

In this case, there are twelve sample moments and seven parameters so we can evaluate the model fit. The null and alternative hypotheses of testing the overall model fit become

$$H_0 : E(m) = \sigma(\tau) \text{ versus } H_1 : E(m) \neq \sigma(\tau) \ . \tag{9}$$

The test statistic is based on the difference between m and $\sigma(\hat{\tau})$ by $F(\hat{\tau})$, T_1, and T_2, where

$$F(\hat{\tau}) = \left\{m - \sigma(\hat{\tau})\right\}^T \hat{V}^{-1} \left\{m - \sigma(\hat{\tau})\right\}, \ T_1 = N \times F(\hat{\tau}), \text{ and } T_2 = T_1 / (1 + F(\hat{\tau})).$$ [15]

Suppose, compared to Model 2, that Model 1 has a smaller chi-square value of the statistic T_2 and does not reject H_0 as shown in equation (9). This implies that Model 1 has better model-data consistency, and for this reason, we consider it the best-fitting model. Therefore, Model 1 reflects the correct causal ordering between variables (*x* causes *y*) (Kano and Shimizu 2003; Shimizu and Kano 2008). The LiNGAM algorithm applies the above test statistics to examine an overall model fit (Shimizu, Hoyer et al. 2006). Noticeably, this method is only feasible when the data is non-Gaussian.

[15] $\hat{V}$ is a weight matrix in GLS estimation that converges in probability to a certain positive definite matrix V.

D. Structural vector autoregressive (SVAR) model

The discussion offered above makes no mention of the time ordering of observations and their possible complications. Generally, such problems are addressed through the use of a vector autoregression (or its derivative error correction representation). Since the vector autoregressive model (VAR model) as proposed by Sims (1980) does not provide enough information to study the causal influence on economic variables in contemporaneous time, the structural vector autoregressive model (SVAR) is used instead. In order to get information in contemporaneous time, we use these machine learning algorithms to infer aspects of the SVAR model on the basis of the statistical distribution of the estimated VAR residuals (Swanson and Granger 1997; Moneta et al. 2013).

Following Hyvärinen et al. (2010), suppose there are *n* related variables at time *t*, $X_t=(X_{1t},...,X_{n,t})$. Define the VAR model as

$$X_t = B_0 X_t + B_1 X_{t-1} + \ldots + B_p X_{t-p} + e_t, \tag{10}$$

where *p* is the number of time lags used and B_0 shows the instantaneous effects and reflects the causal orderings of variables in contemporaneous time. From an SVAR model derive the reduced form VAR model as

$$X_t = \sum_{i=1}^{p} \left(I - B_0\right)^{-1} B_i X_{t-i} + \left(I - B_0\right)^{-1} e_t = \sum_{i=1}^{p} M_i X_{t-i} + \left(I - B_0\right)^{-1} e_t. \tag{11}$$

Calculating the residuals by

$$\hat{u}_t = \left(I - B_0\right)^{-1} e_t = X_t - \sum_{i=1}^{p} \hat{M}_i X_{t-i}, \tag{12}$$

gives

$$\left(I - B_0\right)\hat{u}_t = e_t \Rightarrow \hat{u}_t = B_0 \hat{u}_t + e_t. \tag{13}$$

Graphical-model applications to SVAR identification seek to discover the matrix B_0 of equation (13). When using the PC algorithm, the process starts from tests on conditional independence relations among $\hat{u}_t$. If the error terms are non-normally distributed, perform the LiNGAM algorithm to find the matrix B_0.

IV. Data and results

A. Database description

We use weekly scanner data from Dominick's Finer Foods (DFF), one of the two largest supermarket chains in the Chicago area, courtesy of the "University of Chicago's Kilts Center." The data derives from 88 stores for the period 09/14/89-07/20/94 (253 weeks).[16] DFF prices its products by 16 zones in 4 price tiers: Cub-Fighter, low, medium, and high. DFF's database is an unbalanced panel data. Stores 21, 78, and 101 have the most complete data on sales of Coke Classic and Pepsi-Cola 12-pack and 24-pack, our investigated products. While store 45's data is not as complete, its observations still are over 200. Our purpose of demonstrating differences in these two machine learning algorithms over a modest number of observations (>200) is well-served by this disaggregate analysis. [17] Kadiyali et al. (2000), who aggregate DFF's data across all stores, treat all DFF stores as a common retailer. Villas-Boas (2007), however, defines different retail stores as different retailers, i.e., in her paper store 1 is unique in the metropolitan area, whereas stores 2 and 3 belong to two retail chains. Note that we treat our case as two manufacturers interacting with one retailer because these stores belong to the same retail chain. We assume implicitly that the two manufacturers set a wholesale price for each store. We calculate the CSD manufacturers' selling prices via the provided gross margin measure, a salient characteristic for exploring the relationship between CSD manufacturers and one retailer in a consumer product supply chain. The variables are: (1) retail price (P_r), (2) manufacturer price (P_m), and (3) quantity (Q) number of packs sold. The manufacturers are Coca-Cola Company and PepsiCo.

Generally, the series of the number of packs sold has the highest kurtosis for each store and each product. In most cases, the retail price series has the lowest kurtosis. Moreover, the series of packs sold still has the highest skewness. These

[16] The simulation result from Shimizu, Hyvärinen et al. (2006) indicates that about 80% of causal orderings for three variables can be recovered when the trial number equals 250.

[17] Use of an unbalanced panel specification to identify store-level fixed or random effects would make for an interesting extension; however, it would be beyond the scope of this paper which is to illustrate two machine learning algorithms in a "real world" economic setting.

statistics reveal that the series of Q is far from a normal distribution. The price series positively correlate with each other, whereas there is a negative correlation between price and quantity. The highest correlation is between retail price and manufacturer price, whereas the lowest correlation is between retail price and packs sold in each case.

B. Empirical results

For the LiNGAM algorithm, the prune factor approach is a simplified version of bootstrapping. First, the data are divided into m equally sized groups and then LiNGAM is run on each group to produce a B matrix for each. Then the b_{ij} values are averaged and their standard deviation is calculated. If the absolute value of the mean is less than the prune factor times the standard deviation for that (i, j) entry, the b_{ij} value is set to zero. The number of pieces into which the data are divided is set at 10. Thus, the prune factor indicates the number of standard deviations can be away from the mean bootstrap values. The default value is 1.[18]

We first estimate a reduced-form VAR model (11) and subsequently analyze the estimated residuals on the equation $\hat{u}_t = B_0\hat{u}_t + e_t$. We use the Schwarz Information Criterion (SIC) to choose the optimal time lag for the best multivariate time series fit. We use the raw data of Coke Classic 12-pack for store 45 and store 101 as well as the raw data of Pepsi 12-pack for store 101 directly rather than the residuals, because these series have no lagged effects. Empirically, almost all of the variables series reject the null hypothesis of non-stationarity by using the Augmented Dickey-Fuller test. We verify that the raw data of Coke Classic 12-pack P_r for store 45 and store 101 and Pepsi 12-pack P_r for store 101 series reject the Augmented Dickey-Fuller test with constant drift at a 0.05 significance level. The relevant statistic is shown in Table 3.

[18] This introduction of the prune factor is shown in LiNGAM's MATLAB coding.

Table 3. Test statistic and P-values of the Augmented Dickey-Fuller test for the structural residuals or the raw data

Product	Store	Q	P_r	P_m	Store	Q	P_r	P_m
Coke Classic 12-pack	Store 21	−15.808	−16.801	−18.231	Store 45	−11.909	−1.631	−2.658
		(0.001)	(0.001)	(0.001)		(0.001)	(0.097)	(0.008)
Pepsi 12-pack		−15.526	−16.982	−19.019		−15.819	−15.383	−15.882
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 24-pack		−15.817	−15.547	−18.645		−15.434	−15.396	−17.662
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		−15.993	−15.865	−18.899		−14.963	−14.815	−17.805
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 12-pack	Store 78	−15.876	−15.825	−18.244	Store 101	−12.116	−1.808	−2.699
		(0.001)	(0.001)	(0.001)		(0.001)	(0.067)	(0.007)
Pepsi 12-pack		−15.800	−16.226	−19.135		−11.455	−1.727	−2.448
		(0.001)	(0.001)	(0.001)		(0.001)	(0.08)	(0.014)
Coke Classic 24-pack		−15.922	−15.790	−19.449		−15.796	−15.619	−17.140
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		−15.850	−15.885	−18.887		−15.875	−15.534	−17.727
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)

In addition, we test the structural residuals for non-Gaussianity with the Kolmogorov-Smirnov and Jarque-Bera tests at the 0.05 significance level.[19] The Kolmogorov-Smirnov test yields P-values much smaller than 0.001 for all three variables. In the Jarque-Bera test result, the structural residuals of Coke Classic 12-pack and Pepsi 12-pack P_r for store 21 do not reject the null hypothesis of normal distribution. We note that this is not a problem of the ICA-LiNGAM estimation, i.e., "In the case of just one Gaussian component, we can estimate the model, because the single Gaussian component does not have any other Gaussian components that it could be mixed with" (Hyvärinen et al. 2001: 163). In other words, we can reliably discover a unique correct LiNGAM result when at most one error term is Gaussian (Shimizu, Hoyer et al. 2006). Thus, we conclude that it is appropriate to apply the LiNGAM algorithm. Tables 4 and 5 indicate that the PC algorithm is not sufficient to provide full information of the SVAR models of these variables.

[19] In MATLAB, the null hypothesis of the Kolmogorov-Smirnov test is that the sample in vector x has a standard normal distribution and the null hypothesis of the Jarque-Bera test is that the sample in vector x has a normal distribution with unknown mean and variance.

Table 4. Test statistic and P-values of the Kolmogorov-Smirnov test for the structural residuals or the raw data

Product	Store	Q	P_r	P_m	Store	Q	P_r	P_m
Coke Classic 12-pack	Store 21	0.727	0.222	0.245	Store 45	0.995	0.988	0.904
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 12-pack		0.691	0.226	0.223		0.710	0.296	0.164
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 24-pack		0.760	0.321	0.255		0.702	0.239	0.204
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		0.753	0.272	0.191		0.715	0.182	0.165
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 12-pack	Store 78	0.744	0.309	0.244	Store 101	1.000	0.988	0.904
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 12-pack		0.744	0.290	0.221		1.000	0.992	0.924
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 24-pack		0.731	0.295	0.220		0.774	0.288	0.270
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		0.752	0.243	0.158		0.705	0.271	0.192
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)

Table 5. Test statistic and P-values of the Jarque-Bera test for the structural residuals or the raw data

Product	Store	Q	P_r	P_m	Store	Q	P_r	P_m
Coke Classic 12-pack	Store 21	6470.851	2.545	258.760	Store 45	539.005	35.447	105.209
		(0.001)	(0.234)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 12-pack		4851.141	3.392	68.399		603.817	48.284	90.773
		(0.001)	(0.146)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 24-pack		1043.772	253.373	350.698		279.538	65.192	398.893
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		10129.877	182.178	187.540		1492.669	45.550	187.310
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Coke Classic 12-pack	Store 78	2792.544	40.634	222.903	Store 101	1138.068	19.452	103.201
		(0.001)	(0.001)	(0.001)		(0.001)	(0.002)	(0.001)
Pepsi 12-pack		1942.896	33.694	63.683		1189.658	21.805	117.343
		(0.001)	(0.001)	(0.001)		(0.001)	(0.002)	(0.001)
Coke Classic 24-pack		1926.943	221.088	264.036		1306.867	90.927	424.459
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Pepsi 24-pack		1478.221	146.606	158.030		1996.936	55.076	288.449
		(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)

Figure 1 presents histograms with overlaid Gaussian distributions of the empirical distributions of the structural residuals or the raw data. In general, the histograms lead us to reject the hypothesis of Gaussian distributions. In

most cases, the quantity residuals (or the raw data) are right-skewed (positively skewed), whereas both the retail and manufacturer prices residuals (or the raw data) are slightly left-skewed (negatively skewed). The structural residuals of Coke Classic 12-pack and Pepsi 12-pack P_r for store 21 fail to reject the null hypothesis of normal distribution in the Jarque-Bera test, but successfully reject the null hypothesis of normal distribution in the Kolmogorov-Smirnov test. Based on the graphs of these two variables, we find the peak around the mean and a tendency for the remaining distributions to be symmetrical.

Figures 2 and 3 show that the PC algorithm with 0.1[20] significance level often returns undirected edges. On the other hand, the results of VAR-LiNGAM indicate the pricing pattern, i.e., $P_r \rightarrow P_m$. In other words, DFF has the ability to affect the price charged by these two prominent CSD manufacturers. According to our interpretation, this means that the retailer has greater pricing power.[21] This finding aligns with Kadiyali et al. (2000), who calculated pricing power by studying DFF scanner data for refrigerated juice products in the period 09/14/89-11/25/93, which is similar to our time period. They concluded that even though Tropicana, the brand with higher market share (39.86%), had a higher estimated manufacturer channel profit share than Minute Maid, the brand with lower market share (29.64%), DFF obtained a larger share of total channel profit than both manufacturers in this market, i.e., for Tropicana, DFF received a calculated 58.67% channel profit share and for MinuteMaid, it received a calculated 66.03% channel profit share. We note that the CSD market structure is similar to the yogurt market structure in that two manufacturers, Dannon and General Mills account for almost 62% of the total US yogurt sales. A study by Villas-Boas (2007) of the yogurt market concluded that retailers had greater bargaining power over yogurt manufacturers.[22] The flow is always anticipated and we can see this outcome in the results.[23] When applying the same methodology to other stores' data, $P_r \rightarrow P_m$ or $P_r \rightarrow Q$ is seen often, which implies that different manufacturers undergo a similar pricing pattern.

[20] Spirtes et al. (2001) suggest the proper significance level of the PC algorithm should be 0.1 with sample sizes between 100 and 300.

[21] It is intuitive that the manufacturer with lower market share has less bargaining power. We perform VAR-LiNGAM on store 78's and store 101's Canada Dry 2 liter product that has much lower market share than Coke and Pepsi. Both results showing the causal pattern further support the existence of retailer channel power that dominates manufacturer power.

[22] A reviewer has properly pointed out that the size of the manufacturer or retailer matters. We note, however that our results support other empirical studies applying the structural modeling method, i.e., retailer's channel power dominates the power of manufacturers even when the oligopolistic manufacturers have large market share.

[23] When we reduce the prune factor to 0.5 alternatively, many causal patterns for $P_r \rightarrow Q$, $P_r \rightarrow P_m$, and even $P_m \rightarrow Q \leftarrow P_r$ appear in LiNGAM's outcomes. One possible reason for $P_m \rightarrow Q$ is that the prices charged by CSD manufacturers to retailers usually determine the shelf location for products. The condition, $P_r \rightarrow Q$, is always assumed in most demand analysis and the results show the existence of such a connection.

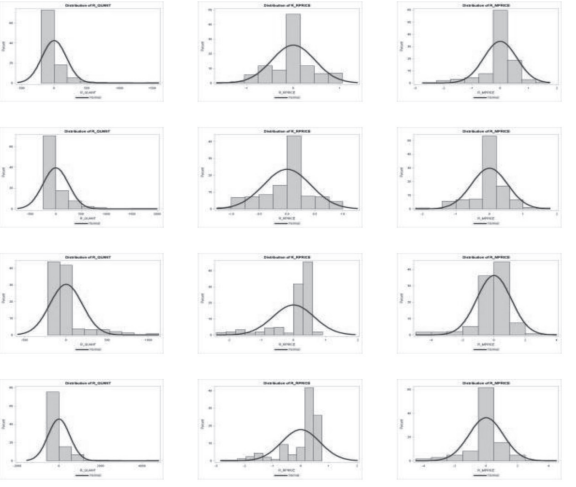
Figure 1. Histograms of the structural residuals or the raw data with overlaid Gaussian distribution with corresponding mean and variance

Coke Classic 12-pack

Pepsi 12-pack

Coke Classic 24-pack

Pepsi 24-pack



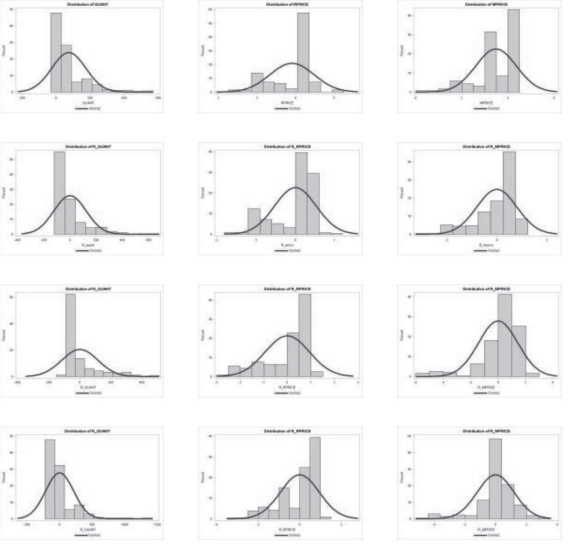
Store 21

Coke Classic 12-pack

Pepsi 12-pack

Coke Classic 24-pack

Pepsi 24-pack



Store 45

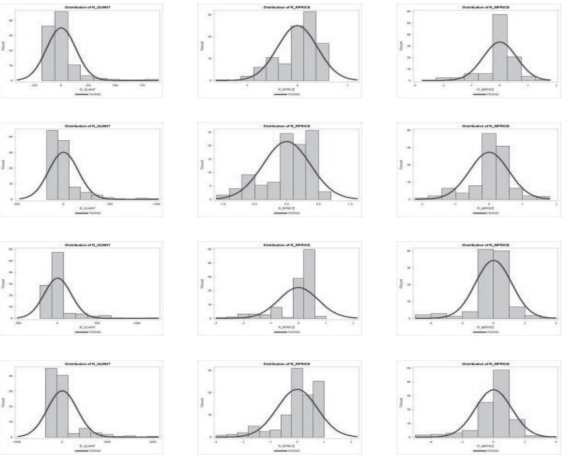
Figure 1. Histograms of the structural residuals or the raw data with overlaid Gaussian distribution with corresponding mean and variance (*continued*)

Coke Classic 12-pack

Pepsi 12-pack

Coke Classic 24-pack

Pepsi 24-pack



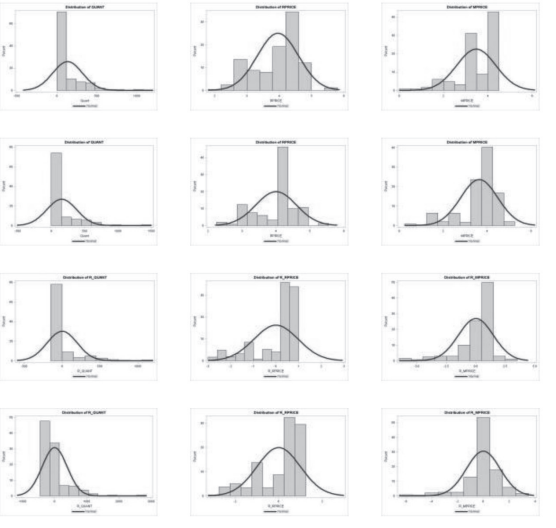
Store 78

Coke Classic 12-pack

Pepsi 12-pack

Coke Classic 24-pack

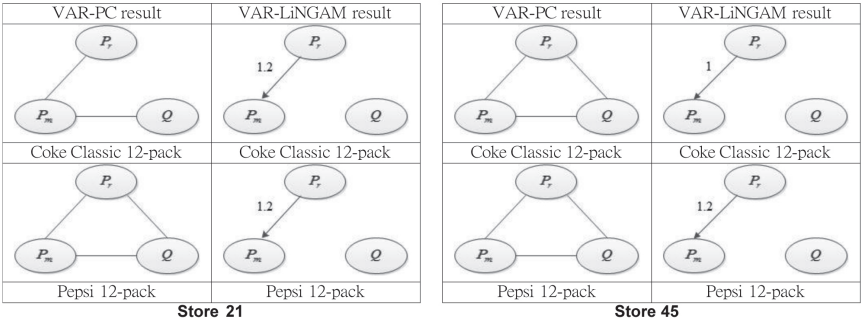
Pepsi 24-pack



Store 101

Note: Column 1 refers to histograms on quantity sold, column 2 to histograms on retail price, column 3 to histograms on manufacturer price.

Figure 2. Empirical graphs of VAR-PC and VAR-LiNGAM estimates for Coke Classic and Pepsi 12-packs



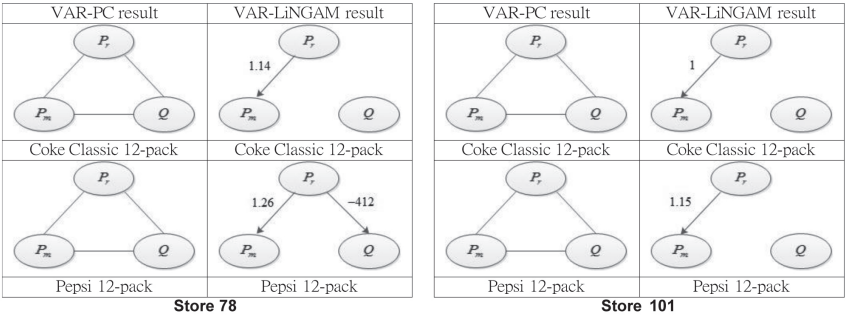


Figure 3. Empirical graphs of VAR-PC and VAR-LiNGAM estimates for Coke Classic and Pepsi 24-packs

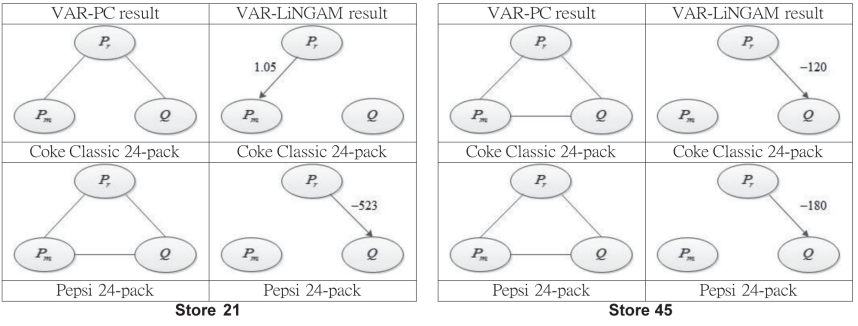
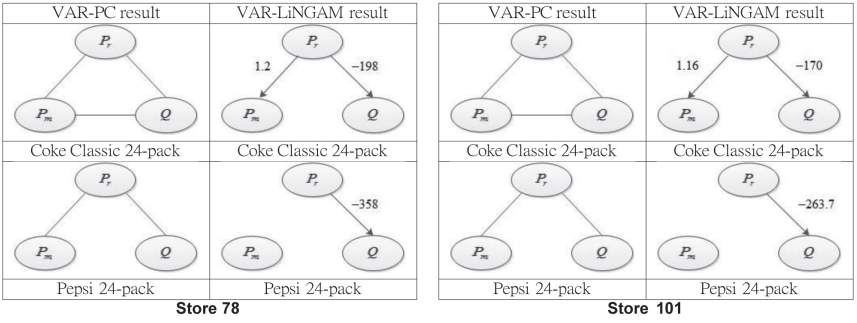


Figure 3. Empirical graphs of VAR-PC and VAR-LiNGAM estimates for Coke Classic and Pepsi 24-packs (*continued*)



V. Conclusions

This paper explored two machine learning algorithms, the PC algorithm and the LiNGAM algorithm, that attempt to uncover causal relationships among the variables, retail price, manufacturer price, and quantity sold, to determine the party with the greater pricing power. Our VAR-LiNGAM graphs indicated clearly that the retail price influenced the prices charged to DFF by the two major carbonated soft drink manufacturers of the drink products under study. By extension, our findings suggest that the retailer's substantial pricing power may derive from its market share in the supermarket industry, its sales of private label products, or from the intense competition between the two CSD manufacturers (Kadiyali et al. 2000; Villas-Boas 2007). Absent the data to estimate inter manufacturer competition, however, we could not verify the latter conjecture. Related research on manufacturer-retailer channel interactions in the soft drink or yogurt category support our findings, e.g., Kadiyali et al. (2000), who estimated a structural model of manufacturer-retailer interactions to determine where pricing power lies in the interaction and concluded that retailers had stronger pricing power as measured by markup than the manufacturer for each national brand of refrigerated juice. Unfortunately, the DFF database lacked information about the manufacturer's margin that could have provided additional verification of our results. Similar to the CSD market, yogurt is produced by a few leading manufacturers. Villas-Boas (2007), who investigated the vertical relationships between yogurt manufacturers

and retailers, found that the retailers had higher bargaining power relative to the yogurt manufacturers. The finding is consistent with our VAR-LiNGAM results.

We also note that Kadiyali et al. (2000) found that the proposed vertical Nash model, the manufacturer Stackelberg model, and the retailer Stackelberg models all were rejected empirically by the Vuong test, which implies that real channel interactions are more complex than what theory assumes. Although our LiNGAM estimation reveals the causal relationship among manufacturer price, retailer price, and sold amount, we caution that it may be inappropriate to conclude that such results can represent a specific pricing game as proposed by theory. At a minimum, the results of LiNGAM show that most retailers have stronger pricing power than CSD manufacturers. Noting that all of the estimated error terms rejected the normality test, the question remains whether the error terms are far enough from the normal distribution to induce a correct estimation in LiNGAM. Hyvärinen et al. (2010) have suggested bootstrapping rather than testing the normality when measuring the accuracy of the estimation. We suggest that future research should examine causal inference under non-Gaussian data.

Finally, this paper provided an overview of an extension of LiNGAM. Lacerda et al. (2008) proposed LiNG-D based on LiNGAM, and added the assumption of stability as an estimation method for cyclic cases that correspond to dynamic systems. When the data are cyclic, and given all (or all but one) non-Gaussian error terms, there is a distribution-equivalent class containing more than one cyclic Structural Equation Model (SEM). Although LiNG-D narrows the class to a distribution-equivalence class of SEMs, it is still possible to have multiple SEMs (Lacerda et al. 2008), in which case the assumption of stability can be used to rule out most of them. Lacerda et al. (2008) provided sufficient conditions for only one SEM in the output of LiNG-D to be stable (the cyclic model to be identifiable): (i) the variables are in equilibrium, i.e., the largest eigenvalue of the coefficient matrix B is smaller than 1 in absolute value; (ii) the cycles are disjoint; and (iii) there are no self-loops.

References

Dominik's Database. Chicago, IL, James M. Kilts Center, Univesity of Chicago Booth School of Business. http://research.chicagobooth.edu/kilts/marketing-databases/dominicks (accessed November 22, 2013).

Drewek, Anna (2010). *A linear non-Gaussian acyclic model for causal discovery*. Master thesis, Swiss Federal Institute of Technology Zurich.

Dodge, Yadolah, and Valentin Rousson (2001). On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician* 55: 51-54.

Gardner, Bruce L. (1975). The farm-retail price spread in a competitive food industry. *American Journal of Agricultural Economics* 57: 399-409.

Haines, Douglas C. (2007). Manufacturer and retailer power in retailer response to trade discounts. *Academy of Marketing Studies Journal* 11: 1-18.

Huang, Dongling, Christian Rojas, and Frank Bass (2008). What happens when demand is estimated with a misspecified model? *The Journal of Industrial Economics* 56: 809-839.

Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja (2001). *Independent component analysis.* New York: John Wiley.

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

Kadiyali, Vrinda, Pradeep Chintagunta, and Naufel Vilcassim (2000). Manufacturer-retailer channel interactions and implications for channel power: An empirical investigation of pricing in a local market. *Marketing Science* 19: 127-148.

Kano, Yutaka, and Shohei Shimizu (2003). Causal inference using nonnormality. In T. Higuchi, Y. Iba, and M. Ishiguro, editors, *Proceedings of the International Symposium on Science of Modeling–the Thirtieth Anniversary of the Information Criterion (AIC), ISM Report on Research and Education 17*, Tokyo, The Institute of Statistical Mathematics.

Kwon, Dae-Heum, and David A. Bessler (2011). Graphical methods, inductive causal inference, and econometrics: A literature review. *Computational Economics* 38: 85-106.

Lacerda, Gustavo, Peter Spirtes, Joseph Ramsey, and Patrik O. Hoyer (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI2008)*, Helsinki.

Moneta, Alessio, Doris Entner, Patrik O. Hoyer, and Alex Coad (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics & Statistics* 75: 705-730.

Pearl, Judea (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Pearl, Judea (2009). *Causality: Models, reasoning, and inference*, 2nd edition. New York, Cambridge University Press

Shimizu, Shohei, Aapo Hyvärinen, Yutaka Kano, and Patrik O. Hoyer (2005). Discovery of non-Gaussian linear causal models using ICA. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-2005),* Quebec.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen and Antti Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7: 2003-2030.

Shimizu, Shohei, Aapo Hyvärinen, Patrik O. Hoyer, and Yutaka Kano (2006). Finding a causal ordering via independent component analysis. *Computational Statistics and Data Analysis* 50: 3278-3293.

Shimizu, Shohei and Yutaka Kano (2008). Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference* 138: 3483-3491.

Shimizu, Shohei and Yoshinobu Kawahara (2010). Non-Gaussian methods for learning linear structural equation models. UAI2010 Tutorial. Osaka University. http://www.ar.sanken.osakau.ac.jp/~sshimizu/papers/ UAI10_ Tutorial_PartI_final_revised.pdf (accessed November 22, 2013).

Shimizu, Shohei, Takanori Inazumiand, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer and Kenneth Bollen (2011). DirectLiNGAM: A Direct method for learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research* 12: 1225-1248.

Sims, Christopher A. (1980), Macroeconomics and reality. *Econometrica* 48: 1-48.

Spirtes, Peter, Clark Glymour, and Richard Scheines (2001). *Causation, prediction, and search*, 2nd edition. Cambridge, MA, MIT Press.

Spirtes, Peter, Richard Scheines, Clark Glymour, Thomas Richardson, and Christopher Meek (2004). Causal inference. *Handbook of quantitative methodology in the social sciences*.

Spirtes, Peter (2005). Graphical models, causal inference, and econometric models. *Journal of Economic Methodology* 12: 3-34.

Spirtes, Peter (2010). Introduction to causal inference. *The Journal of Machine Learning Research* 99: 1643-1662.

Stone, James V. (2004). *Independent component analysis: A tutorial introduction. Cambridge*, MA, MIT Press.

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

Villas-Boas, Sofia Berto (2007). Vertical relationships between manufacturers and retailers: Inference with limited data. *Review of Economics Studies* 74: 625-652.