

**CAUSAL CONNECTION SEARCH AND STRUCTURAL DEMAND  
MODELING ON RETAIL-LEVEL SCANNER DATA**

A Dissertation

by

PEI-CHUN LAI

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Agricultural Economics

**CAUSAL CONNECTION SEARCH AND STRUCTURAL DEMAND  
MODELING ON RETAIL-LEVEL SCANNER DATA**

A Dissertation

by

PEI-CHUN LAI

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee, David Bessler  
Yanyuan Ma

Committee Members, H. Alan Love  
Ximing Wu

Head of Department, Steven Puller  
John P. Nichols

December 2010

Major Subject: Agricultural Economics

**ABSTRACT**

Causal Connection Search and Structural Demand Modeling  
on Retail-Level Scanner Data. (December 2010)

Pei-Chun Lai, B.A., National Taiwan University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. David Bessler  
Dr. Yanyuan Ma

Many researchers would be interested in one question: If a change of  $X$  is made, will  $Y$  be influenced in response? However, while a lot of statistical methods are developed to analyze association between variables, how to find a causal relationship among variables is relatively neglected.

The PC algorithm, developed on the basis of Pearl, Sprites, Glymour, and Scheines's studies, is used to find the causal pattern of the real-world observed data. However, PC in Tetrad produces a class of directed acyclic graphs (DAGs) that are statistically equivalent under a normal distribution, and therefore such a distributional assumption causes a series of unidentifiable DAGs because of the same joint probability.

In 2006 Shimizu, Hoyer, Hyvärinen, and Kerminen developed the Linear Independent Non-Gaussian Model (LiNGAM) to do a causal search based on the independently non-Gaussian distributed disturbances by applying higher-order moment structures. The research objective of this dissertation is to examine whether the LiNGAM is helpful relative to the PC algorithm, to detect the causal relation of non-

normal data. The LiNGAM algorithm is implemented by first doing independent component analysis (ICA) estimation and then discovering the correct ordering of variables. Thus, the procedures of ICA estimation and the process of finding the correct causal orderings in LiNGAM are illustrated. Next, we do a causal search on the retail-level scanner data to investigate the pricing interaction between the manufacturer and the retailer by applying these two algorithms. While PC generates the set of indistinguishable DAGs, LiNGAM gives more exact causal patterns. This work demonstrates the algorithm based on the non-normal distribution assumption makes causal associations clearer. In Chapter IV, we apply a classical structural demand model to investigate the consumer purchase behavior in the carbonated soft drink market. Unfortunately, when further restrictions are imposed, we cannot get reasonable results as most researchers require. LiNGAM is applied to prove the existence of endogeneity for the brand's retail price and verify that the brand's wholesale price is not a proper instrument for its retail price. Therefore, consistent estimates cannot be derived as the theories suggest. These results imply that economic theory is not always found in restriction applied to observational data.

**DEDICATION**

*To my dearest father,*

*Fu-Tsung Lai*

## ACKNOWLEDGEMENTS

First, I am extremely thankful for my advisor, Dr. David Bessler, who always gives me confidence and guidance in the direction of my research life. He is the most humble, patient, and kind person I have met. I especially admire his warm attitude to every person around him and his strict academic discipline. I hope I can be as good a teacher as him in my future career life.

I also want to thank to all of my other committee members: Dr. Yanyuan Ma, Dr. Alan Love, Dr. Ximing Wu, and Dr. Steven Puller. They provide me with substantial insight of econometric methods and industrial organization theory. I have a great experience when attending their classes and enjoy a lot of personal conversation with them. They are all wonderful teachers. Moreover, I am grateful for all my classmates and department friends; Francisco Fraire, Wei Huang, Yongxia Cai, Justus, Seongwoo Kim, Yuan Yan, Thein, Memory, Aklesso, Amy, Meng-Shiuh Chang, Siyi Feng and WeiWei Wang, etc. Without their help, friendship, and encouragement, I could not have finished my PhD studies.

Finally, it would be impossible to have my research career without my family's love and support. I am especially grateful to my father and my fiancé. My father always accompanies me in every important step of my life and provides me constant trust, love and consistent support. My fiancé, TengXi Wang, gives a lot of comfort and takes care of me well especially during the tough time of writing this dissertation. Both of them

shared most of my frustrating moments and supported me to finish this dissertation. This dissertation is dedicated to them!

To all of you, thank you!

**NOMENCLATURE**

AIDS	Almost Ideal Demand System
CF	Caffeine-Free
CMC	Causal Markov Condition
CSD	Carbonated Soft Drink
DAG	Directed Acyclic Graph
DFE	Dominick's Finer Foods
DM	Distance Metric
DPSU	Dr Pepper/Seven-Up Companies, Inc
ICA	Independent Component Analysis
LA/AIDS	Linear Approximate Almost Ideal Demand System Model
LiNGAM	Linear Independent Non-Gaussian Model
LSEM	Linear Structural Equation Modeling
OLS	Ordinary Least Squares
SVAR	Structural Vector Autoregressive Models
VAR	Vector Autoregressive Models



## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
DEDICATION .....	v
ACKNOWLEDGEMENTS .....	vi
NOMENCLATURE.....	viii
TABLE OF CONTENTS .....	ix
LIST OF FIGURES.....	xii
LIST OF TABLES .....	xiii
 CHAPTER	
I INTRODUCTION.....	1
Causation and Linear Structural Equation Modeling .....	1
Causation .....	1
Linear Structural Equation Modeling.....	2
Research Objective.....	3
II CAUSAL MODEL IDENTIFICATION AND INDEPENDENT COMPONENT ANALYSIS .....	5
Introduction .....	5
Graph Theory .....	5
Paths and Edge Sequences .....	5
Directed Acyclic Graph.....	6
Causal Markov Condition .....	7
D-separation .....	7
Detect a Causal Ordering between Two Variables .....	8
Independent Component Analysis .....	20
Basic Concept of Independent Component Analysis .....	21
Introduction of Entropy .....	22
Entropy of Continuous Variables .....	23

CHAPTER	Page
	24
	24
	29
	29
	30
<b>III</b>	
<b>CAUSAL SEARCH ON THE PRICING LEADERSHIP BETWEEN THE MANUFACTURER AND THE RETAILER .....</b>	<b>32</b>
Introduction .....	32
Search Algorithm for Finding Causal Structure.....	32
PC Algorithm .....	33
Linear Non-Gaussian Acyclic Models .....	35
LiNGAM Discovery Algorithm .....	37
Method to Prune the Edges .....	42
Determine the Direction of Causality.....	46
Structural Vector Autoregressive Model with LiNGAM.....	49
Strategic Interaction between Firms.....	50
Data and Empirical Result.....	52
Database Description .....	52
Characteristics of Data .....	53
Prune Factor .....	54
Empirical Results .....	55
Conclusion.....	65
<b>IV</b>	
<b>STRUCTURAL DEMAND MODEL FOR THE U.S. CARBONATED SOFT DRINK MARKET .....</b>	<b>67</b>
Introduction .....	67
The Carbonated Soft Drink Industry in the United States.....	69
Quantitative Methods .....	71
The Demand Model.....	71
Distance Metric Approach .....	74
Household Demographics .....	77
Data and Preliminary Data Statistics.....	77
Scanner Data .....	77
Data Preparation .....	78
Distance Metrics.....	81
Continuous Distance Measures with Continuous Brand Attributes .....	82

CHAPTER	Page
Discrete Distance Measures with Continuous Brand Attributes .....	85
Discrete Distance Measures with Discrete Brand Attributes .....	87
Estimated Results .....	87
Preliminary OLS Regression Result of Disaggregation .....	88
Preliminary OLS Regression Result after Aggregation .....	89
Empirical Problem and Conclusion.....	89
 V CONCLUSIONS .....	 94
Key Findings .....	94
Possible Future Research .....	96
 REFERENCES.....	 97
 APPENDIX A .....	 102
 APPENDIX B .....	 109
 APPENDIX C .....	 113
 APPENDIX D .....	 117
 APPENDIX E.....	 120
 APPENDIX F .....	 123
 VITA .....	 128

**LIST OF FIGURES**

FIGURE	Page
1 A Directed Acyclic Graph.....	6
2 The Corresponding Causal Connection of Equation (82).....	38
3 The Corresponding Causal Association of Equation (88).....	41
4 Causal Connection between the Expenditure Share and First Difference of the Retail Prices Searched by LiNGAM with Prune Factor=0.7 .....	92
5 Causal Connection between the Retail Price and the Wholesale Price Searched by LiNGAM with Prune Factor=0.7 .....	92

## LIST OF TABLES

TABLE		Page
1	Empirical Graphs of LiNGAM and PC Estimates for CSDs .....	57
2	Causal Associations of Residuals from VAR-LiNGAM and VAR-PC Estimates for CSDs .....	61
3	Empirical Graphs of LiNGAM and PC Estimates for CSDs with Lower Prune Factor and Significance Level.....	63
4	General Traits of Typical Household in Demographic Cluster in DFF Database .....	80
5	Statistics of Demographics for Store Cluster in DFF Database .....	83
6	OLS Regression Results of Estimated Coefficient on Distance Metrics after Aggregation.....	90
F1	OLS Results of Estimated Coefficient on Distance Metrics of Disaggregate CSD (Package Size is Treated as a Continuous Variable) ..	123
F2	OLS Results of Estimated Coefficient on Distance Metrics of Disaggregate CSD (Package Size is Treated as a Dummy Variable) .....	124
F3	The Comparison between Our Disaggregate Estimation Results and Dube's Results .....	125
F4	List of Aggregate Brands with the Average Retail Price and Their Shares of All Sold Volume.....	126
F5	Attributes of CSD Brands Used in Our Dataset.....	127

## CHAPTER I

### INTRODUCTION

While there are plenty of statistical methods available to analyze correlation between variables, the question of a causal relationship among uncontrolled variables is relatively neglected (Dodge and Rousson, 2001). A typical procedure used in study of the structural equation model is, to collect data, establish a particular model a priori according to the researchers' belief, test the model fit, and evaluate whether the association among variables is significant (Glymour, 2010). However, the causal inference analysis is also quite important and, for example, a causal understanding of the data is helpful to predict the possible results of given interventions or policies (Shimizu, Hyvärinen, Kano and Hoyer, 2005). Noticeably, even if the association between  $X$  and  $Y$  is statistically significant, it does not imply that  $X$  must be the cause of  $Y$  or  $Y$  must be the cause of  $X$  (Miller, 2005). Causal relation can be detected by doing controlled experiment, but real-world environment is always uncontrolled, so it is necessary to use other advanced methodologies to search the causal relation of real-world data.

### Causation and Linear Structural Equation Modeling

#### Causation

It is oftentimes assumed that:

1. If  $X$  is a cause of  $Y$ , and  $Y$  is a cause of  $Z$ , then  $X$  is a cause of  $Z$  (transitivity),

---

This dissertation follows the style of *RAND Journal of Economics*.

2. Variable  $X$  cannot cause itself (irreflexive), and
3.  $X$  is a cause of  $Y$  does not mean  $Y$  is a cause of  $X$  (not symmetric).

When detecting the causal relationship of variables, actually, we are asking which variable occurred first.

### **Linear Structural Equation Modeling**

The general equation form of causation is always shown that:

$$x_i = f(pa(x_i), u_i), \quad i = 1, \dots, n, \quad (1)$$

where  $pa(x_i)$ , Markovian Parents, represents the set of variables that determine the value of  $x_i$  and the  $u_i$  symbolizes an error term. Equation (1) can be written as a generalization of the linear structural equation model (LSEM)

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, \dots, n, \quad (2)$$

where  $pa(x_i)$  corresponds to those  $x_k$  s. Applying the causal interpretation, the variables on the right hand side,  $x_k$  s, of an LSEM are the directed causes of the dependent variable  $x_i$ . Moreover, if  $x_2$  is a direct cause of  $x_1$ , then there is an edge from  $x_2$  to  $x_1$  in the corresponding path diagram ( $x_2 \rightarrow x_1$ ) (Pearl, 2009). However, establishing a proper structural model always depends on the researchers' belief or follows prior theory, but sometimes personal knowledge or theory may not be consistent with the reality. Probabilistic methods based on Pearl (2009), Glymour, Scheines and Spirtes' research (2001) are widely used to examine causation among uncontrolled (or observed) variables. In their methodology, mean and covariance matrix are commonly applied, based on the assumption of data's normal distribution under the central limit theorem. However, this

is usually not enough. In the chapters that follow, we explore conditions to infer causal ordering and study scanner data refer to soft drink market as a case study.

### **Research Objective**

The presumption of a normal distribution for sample data causes many unidentifiable causal patterns because those graphs have the same joint probability. Moreover, non-normal distributed data are oftentimes found in empirical applications. Shimizu et al. (2006) developed Linear Independent Non-Gaussian Model (LiNGAM) to do causal search based on the independently non-Gaussian distributed disturbances by applying higher-order moment structures. Therefore, whether higher-order moment structures are beneficial in detecting causal relation of variables with non-normal probability distribution, relative to the normal distribution presumption, is a main issue discussed in this thesis. In Chapter II, initially, the basic element of graph theory is introduced. Moreover, the procedures of searching the causal connection between two variables by using skewness, kurtosis, higher-order covariance and correlation are explained. The LiNGAM algorithm is implemented by first doing independent component analysis (ICA) estimation and then discovering the correct ordering of variables. Therefore, the theorem behind ICA and the steps of finding demixing matrix by doing ICA is introduced at the end of Chapter II. At the beginning of Chapter III, the comparison of PC algorithm, worked reliably on normally distributed or symmetrically non-normal distributed data, and LiNGAM algorithm, works better on the more non-Gaussian data, is made. Then the process of deriving the correct causal orderings in



LiNGAM is illustrated. Finally, we do causal search on the retail-level data to investigate the pricing power between manufacturer and retailer in a distribution channel by applying these two algorithms. While PC generates a set of indistinguishable graph structures with the same joint probability, LiNGAM gives more exact causal patterns. This demonstrates the algorithm based on the non-normal distribution assumption make causal relations clearer. Unlike other chapters, we apply classical structural demand model to investigate the consumer purchase behavior in carbonated soft drink market. Unfortunately, when further restrictions are imposed as the theory suggests, we do not obtain reasonable results as most researchers require. This implies that economic theory may not correspond to the movement of real-world data; it also shows the importance of causal inference research. At the end of Chapter IV, the graphical causal inference method can be used to test the appropriateness of a possible instrument.

## CHAPTER II

### CAUSAL MODEL IDENTIFICATION AND INDEPENDENT COMPONENT ANALYSIS

#### Introduction

One of the main targets in social science is the discovery of causal structure among variables. Directed graphs, based on the research works of Pearl (2009); Sprites, Glymour and Scheines (2001), can sometimes be used to find the actual data-based causal relationships. In these studies, I briefly introduce the relevant concepts of graphical representations, related algorithms, and discuss their empirical applications.

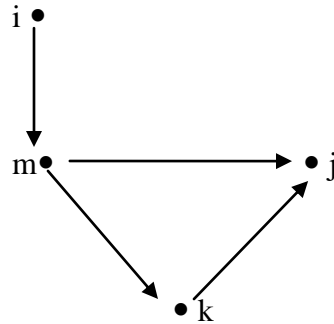
#### Graph Theory<sup>1</sup>

##### Paths and Edge Sequences

A directed graph  $G = (V, E)$  consists of a set  $V$  of vertices (or objects) and a set  $E$  of edges (or links) that connect arbitrary pairs of elements of  $V$ . Two vertices are said to be adjacent if they are connected by the same edge. A path in a graph is a sequence of consecutive edges (Pearl, 2009). Suppose all edges are directed as Figure 1 shows, and then we have a directed graph. A directed path is that, given an sequence of distinct vertices  $(V_1, \dots, V_n)$ ,  $V_i$  is a direct cause of  $V_{i+1}$  for all  $0 \leq i \leq n-1$ . Therefore, the path  $((i, m), (m, k))$  is directed, but the path  $((i, m), (m, j), (j, k))$  is not (Meek, 1995; Zhang, 2008a).

---

<sup>1</sup> Parts of this section are summarized from Pearl's book (2009) and Glymour's paper (2010). Especially, the definitions of causal Markov condition and d-separation come directly from their texts.

**FIGURE 1 A Directed Acyclic Graph****Directed Acyclic Graph**

A directed cycle is a directed path that starts and ends in the same vertex. A directed acyclic graph (DAG) is a directed graph without directed cycles. The vertices of the graph represent the variables. Suppose there is a directed path starting in  $X$  and ending in  $Y$ , and then it is said that  $X$  is an ancestor of  $Y$ , and  $Y$  is a descendant of  $X$ .

Other relationships between  $X$  and  $Y$  in a graph  $G$  can be summarized as

$$\text{If } \begin{cases} Y \leftrightarrow X \\ Y \rightarrow X \\ Y \leftarrow X \end{cases} \text{ in } G \text{ then } Y \text{ is a } \begin{cases} \text{spouse} \\ \text{parent} \\ \text{child} \end{cases} \text{ of } X, \text{ and } \begin{cases} Y \in SPO_G(X) \\ Y \in PAR_G(X), \\ Y \in CHI_G(X) \end{cases}$$

and the set of  $X$ 's descendants is denoted  $DES_G(X)$  (Zhang, 2008b; Richardson and Spirtes, 2002).

$Z$  is a collider if  $Z$  is at the head of two edges on a path. If a path contains two edges having common head  $Z$ , and the respective tails  $X$  and  $Y$  are not adjacent, then  $Z$  is called an unshielded collider (Glymour, 2010).

## Causal Markov Condition

The Causal Markov Condition (CMC), defined by Spirtes, Glymour, and Scheines (2001), describes that: Suppose  $P$  is a probability distribution over the vertices in  $V$  of a given directed acyclic graph  $G$ . Then  $G$  and  $P$  satisfy the Causal Markov Condition if and only if for every  $X$  in the set  $V$ ,  $X$  is independent of  $V \setminus (DES_G(X) \cup PAR_G(X))$  conditional on  $PAR_G(X)$ .<sup>2</sup> In other words, the Causal Markov Condition states that any vertex (or variable) in a DAG  $G$  is independent of its nondescendants conditional on its parents (Glymour, 2010). Then the joint probability density function over all the vertices  $f(V)$  satisfying the CMC is given by

$$f(V) = \prod_{v \in V} f(v | PAR_G(v)) \quad (3)$$

## D-separation

The CMC provides a connection between graphical models and the joint probability over the corresponding variables. It does not, however, directly provide a computable procedure for determining independence relations among variables. Pearl developed d-separation which indicates the conditional and marginal probabilistic independence entailed by the CMC (Pearl, 1988; Glymour, 2010).

For all variables  $X, Y$  where  $X \neq Y$  in  $G$  and subsets  $\mathbf{Z}$  not containing  $X$  and  $Y$ ,  $X$  and  $Y$  are d-separated given  $\mathbf{Z}$  if, and only if, every path from  $X$  to  $Y$  contains at least one variable  $Z_i$  such that either:

1.  $Z_i$  is a collider, and no descendant of  $Z_i$  (including  $Z_i$  itself) is in  $\mathbf{Z}$ ; or

---

<sup>2</sup>  $X \perp\!\!\!\perp Y$  denote the relative complement of  $Y$  with respect to  $X$ .

2.  $Z_i$  is not a collider, and  $Z_i$  is in  $\mathbf{Z}$ .

For example, the causal chains  $X \rightarrow Z \rightarrow Y$  and causal forks  $X \leftarrow Z \rightarrow Y$  represent cases where  $Z$  d-separates  $X$  and  $Y$ . That is,  $X$  and  $Y$  are probabilistically dependent; but conditioning on  $Z$ ,  $X$  and  $Y$  are independent, ( $P(X, Y) \neq P(X)P(Y)$  but  $P(X, Y|Z) = P(X|Z)P(Y|Z)$ ). For the case of the causal chain, it means that the direct cause,  $Z$ , removes the effect from indirect cause,  $X$ , on  $Y$ . In contrast, unshielded colliders (inverted fork),  $X \rightarrow Z \leftarrow Y$ , represent the cases where two causes have a common effect. Knowledge of such colliders tends to make the causes dependent. Information about the collider ( $Z$ ) and the occurrence of one of the causes makes the other more or less possible. That is,  $X$  and  $Y$  are probabilistically independent conditional on the set of variables excluding  $Z$ , but are probabilistically dependent conditional on  $Z$  ( $P(X, Y) = P(X)P(Y)$  but if we condition on  $Z$ ,  $P(X, Y|Z) \neq P(X|Z)P(Y|Z)$ ) (Pearl, 2009).

### Detect a Causal Ordering between Two Variables<sup>3</sup>

When we evaluate a linear regression model of  $X$  and  $Y$ , we may hesitate between the equation

$$Y = \beta X + \varepsilon_Y \quad (4)$$

where  $X$  is independent of  $\varepsilon_Y$ , and the equation

---

<sup>3</sup> Parts of this section follow Dodge and Rousson's paper (2001) and Hoover's article (2009b). Particularly, the operations of skewness and correlation between variables come directly from Dodge and Rousson's text.

$$X = \eta Y + \varepsilon_X \quad (5)$$

where  $Y$  is independent of  $\varepsilon_X$  and  $\varepsilon_X$  and  $\varepsilon_Y$  are disturbances. These coefficients  $\alpha, \beta, \delta$  and  $\eta$  are fixed. From the viewpoint of causal structure,  $X$  causes  $Y$  ( $X \rightarrow Y$ ) is an implication behind the equation (4) while  $Y$  causes  $X$  ( $Y \rightarrow X$ ) is an implication behind the equation (5).

When  $X$  and  $Y$  are both normally distributed, equation (4) and equation (5) are observationally equivalent; in other words, these two equations are not identified. The reason is that: If  $X \rightarrow Y$ , then we can write

$$\begin{aligned} X &= \varepsilon_X \\ Y &= \beta X + \varepsilon_Y = \beta \varepsilon_X + \varepsilon_Y \end{aligned} \quad (6)$$

where each error term is distributed normally with mean zero and variances  $\sigma_X^2$  and  $\sigma_Y^2$ . Thus,  $X$  is normally distributed with  $N(0, \sigma_X^2)$ ;  $Y$  is normally distributed with  $N(0, \beta^2 \sigma_X^2 + \sigma_Y^2)$  and  $\text{cov}(X, Y) = \beta \sigma_X^2$ . Instead, if  $Y \rightarrow X$ , then

$$\begin{aligned} X &= \eta Y + \varepsilon_X = \eta \varepsilon_Y + \varepsilon_X \\ Y &= \varepsilon_Y \end{aligned} \quad (7)$$

Therefore,  $X$  is normally distributed with  $N(0, \eta \sigma_Y^2 + \sigma_X^2)$ ;  $Y$  is normally distributed with  $N(0, \sigma_Y^2)$  and  $\text{cov}(X, Y) = \eta \sigma_Y^2$ . Obviously, although these two cases are different in causal relations, they have indistinguishable mean, variance, and covariance structures. Hence, we cannot tell equation (4) from (5). When two equations are observationally equivalent, the only way to estimate the parameters is to presume a specific causal structure in advance (Hoover, 2009b).

However, if both  $X$  and  $Y$  have non-normal distributions, then equation (4) and equation (5) are distinguishable by using third- and even fourth-order correlation structure of variables. The following argument is taken from Dodge and Rousson (2001):

The expected value of a random variable  $g(X)$  is denoted by  $E[g(X)]$  and the  $n$ th moment of  $X$ ,  $\mu'_n$ , is defined as

$$\mu'_n(X) = E(X^n) \quad (8)$$

The  $n$ th central moment of  $X$ ,  $\mu_n$ , is defined as

$$\mu_n(X) = E(X - \mu_X)^n = E(X - \mu'_1(X))^n \quad (9)$$

The Taylor expansion series of  $f(X)$  about origin is given as:

$$f(X) = f(0) + f'(0)X + \frac{f''(0)}{2!}X^2 + \dots + \frac{f^{(n)}(0)}{n!}X^n + \dots, \quad (10)$$

Suppose  $f(X) = e^{tX}$ , and thus  $f'(X) = t \cdot e^{tX}$ ,  $f''(X) = t^2 \cdot e^{tX}$ , ...,  $f^{(n)}(X) = t^n \cdot e^{tX}$

Therefore, the moment generating function of  $X$ , denoted by  $M_X(t)$ , is specified by apply equation (10):

$$\begin{aligned} M_X(t) &= E(e^{tX}) = E\left(1 + tX + \frac{t^2}{2!}X^2 + \dots + \frac{t^n}{n!}X^n + \dots\right) = 1 + tE(X) + \dots + \frac{t^n}{n!}E(X^n) + \dots \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} \mu'_n \end{aligned} \quad (11)$$

The cumulants  $\kappa_n(X)$  of a random variable  $X$  are defined by the cumulant generating function, the natural logarithm of the moment-generating function:

$$g_X(t) = \ln(E(e^{tX})) = \sum_{n=1}^{\infty} \kappa_n(X) \frac{t^n}{n!} = \kappa_1(X)t + \kappa_2(X) \frac{t^2}{2!} + \kappa_3(X) \frac{t^3}{3!} + \dots \quad (12)$$

The cumulants are then given by derivatives (at zero) of  $g_X(t)$  with respect to  $t$ :

$$\begin{aligned} \kappa_1(X) &= g_X'(0), \\ \kappa_2(X) &= g_X''(0), \\ &\vdots \\ \kappa_n(X) &= g_X^{(n)}(0). \end{aligned} \quad (13)$$

Suppose the true model is:

$$Y = \alpha + \beta X + \varepsilon, \quad (14)$$

where  $\alpha$  and  $\beta$  are constants. Then we have

$$\rho_{XY} = \beta \frac{\sigma_X}{\sigma_Y} \quad (15)$$

Thus, the cumulant generating function on  $Y$  is given as:

$$\begin{aligned} g_Y(t) &= \ln(E(e^{tY})) = \sum_{n=1}^{\infty} \kappa_n(Y) \frac{t^n}{n!} = \kappa_1(Y)t + \kappa_2(Y) \frac{t^2}{2!} + \kappa_3(Y) \frac{t^3}{3!} + \dots \\ &= \ln(E(e^{t(\alpha + \beta X + \varepsilon)})) = \ln(E(e^{t\alpha}) \cdot E(e^{t\beta X}) \cdot E(e^{t\varepsilon})) = \ln(E(e^{t\alpha})) + \ln(E(e^{t\beta X})) + \ln(E(e^{t\varepsilon})) \\ &= \ln(e^{t\alpha}) + \ln(E(e^{t\beta X})) + \ln(E(e^{t\varepsilon})) \\ &= t\alpha + \left[ \kappa_1(X)t\beta + \kappa_2(X) \frac{t^2 \beta^2}{2!} + \dots + \kappa_n(X) \frac{t^n \beta^n}{n!} \right] + \left[ \kappa_1(\varepsilon)t + \kappa_2(\varepsilon) \frac{t^2}{2!} + \dots + \kappa_n(\varepsilon) \frac{t^n}{n!} \right] \end{aligned} \quad (16)$$

$$\begin{aligned} \kappa_1(Y) &= g_Y'(0) = \alpha + \beta \kappa_1(X) + \kappa_1(\varepsilon) \\ \kappa_2(Y) &= g_Y''(0) = \beta^2 \kappa_2(X) + \kappa_2(\varepsilon) \end{aligned}$$

and thus for  $n \geq 3$ ,

$$\kappa_n(Y) = g_Y^{(n)}(0) = \beta^n \kappa_n(X) + \kappa_n(\varepsilon) \quad (17)$$



Setting  $f(X) = \ln(1 + X)$  and apply this function into the formula of Taylor expansion as shown in (10). Then the fact is derived as:

$$\ln(1 + z) = \sum_{j=1}^{\infty} \left( (-1)^{j-1} \frac{z^j}{j} \right) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots, \quad (18)$$

Suppose  $z = tE(X) + \dots + \frac{t^n}{n!} E(X^n) + \dots$ . If we only consider the polynomial of degree 4

of the cumulant generating function, then

$$\begin{aligned} g_X(t) &= \ln(E(e^{tX})) = \ln(1 + z) \\ &\approx tE(X) + \frac{t^2}{2!} E(X^2) + \frac{t^3}{3!} E(X^3) + \frac{t^4}{4!} E(X^4) \\ &\quad - \frac{t^2 [E(X)]^2}{2} - \frac{t^3 E(X)E(X^2)}{2} - \frac{t^4 [E(X^2)]^2}{8} - \frac{t^4 E(X)E(X^3)}{3!} + \frac{t^3 [E(X)]^3}{3} \\ &\quad + \frac{t^4 [E(X)]^2 E(X^2)}{2} - \frac{t^4 [E(X)]^4}{4} \end{aligned} \quad (19)$$

Therefore, the first, second, third, and fourth cumulants of  $X$  are

$$\begin{aligned} \kappa_1(X) &= g_X'(0) = E(X) = \mu_1(X), \\ \kappa_2(X) &= g_X''(0) = E(X^2) - [E(X)]^2 = \mu_2(X), \\ \kappa_3(X) &= g_X'''(0) = E(X^3) - 3E(X)E(X^2) + 2[E(X)]^3 = \mu_3(X), \\ \kappa_4(X) &= g_X^{(4)}(0) = E(X^4) - 3[E(X^2)]^2 - 4E(X)E(X^3) \\ &\quad + 12[E(X)]^2 E(X^2) - 6[E(X)]^4 = \mu_4(X) - 3[\kappa_2(X)]^2. \end{aligned} \quad (20)$$

Obviously, the second and third central moments are respectively equal to second and third cumulants.

For  $n > 2$ , the  $n$ th standardized cumulants of  $X$  are defined as

$$\gamma_n(X) = \frac{\kappa_n(X)}{[\kappa_2(X)]^{n/2}} = \frac{\kappa_n(X)}{\sigma_X^n}, \quad n = 3, 4, \dots \quad (21)$$

$$\begin{aligned}
\gamma_n(\alpha + \beta X) &= \frac{\kappa_n(\alpha + \beta X)}{[\kappa_2(\alpha + \beta X)]^{n/2}} = \frac{\beta^n \cdot \kappa_n(X)}{[\beta^2 \kappa_2(X)]^{n/2}} = \frac{\beta^n \cdot \kappa_n(X)}{\beta^n \cdot [\kappa_2(X)]^{n/2}} \\
&= \frac{\kappa_n(X)}{[\kappa_2(X)]^{n/2}} = \gamma_n(X)
\end{aligned} \tag{22}$$

It means the standardized cumulant are invariant under translations and scaling so the standardized cumulant can be used to judge the shape of the probability density.

The error term,  $\varepsilon$ , is assumed to be normally distributed with mean  $\mu_\varepsilon$  and variance  $\sigma_\varepsilon^2$ , denoted by  $\varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$ .

For the normally distributed  $\varepsilon$ , the moment generating function is

$$\begin{aligned}
M_\varepsilon(t) &= E(e^{t\varepsilon}) = \int_{-\infty}^{\infty} e^{t\varepsilon} f_\varepsilon(\varepsilon) d\varepsilon \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\varepsilon-\mu_\varepsilon)^2}{2\sigma_\varepsilon^2}} e^{t\varepsilon} d\varepsilon = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{\varepsilon^2 - 2\varepsilon\mu_\varepsilon + \mu_\varepsilon^2 - 2\sigma_\varepsilon^2 t\varepsilon}{2\sigma_\varepsilon^2}\right]} d\varepsilon \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{\varepsilon^2 - 2(\mu_\varepsilon + \sigma_\varepsilon^2 t)\varepsilon + (\mu_\varepsilon + \sigma_\varepsilon^2 t)^2}{2\sigma_\varepsilon^2}\right]} e^{\left(\mu_\varepsilon t + \frac{\sigma_\varepsilon^2 t^2}{2}\right)} d\varepsilon \\
&= e^{\left(\mu_\varepsilon t + \frac{\sigma_\varepsilon^2 t^2}{2}\right)} \cdot \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{\varepsilon^2 - 2(\mu_\varepsilon + \sigma_\varepsilon^2 t)\varepsilon + (\mu_\varepsilon + \sigma_\varepsilon^2 t)^2}{2\sigma_\varepsilon^2}\right]} d\varepsilon \\
&= e^{\left(\mu_\varepsilon t + \frac{\sigma_\varepsilon^2 t^2}{2}\right)} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{[\varepsilon - (\mu_\varepsilon + \sigma_\varepsilon^2 t)]^2}{2\sigma_\varepsilon^2}\right]} d\varepsilon = e^{\left(\mu_\varepsilon t + \frac{\sigma_\varepsilon^2 t^2}{2}\right)}
\end{aligned} \tag{23}$$

Consequently, the cumulant generating function of normal distributed  $\varepsilon$  is

$$g_\varepsilon(t) = \ln(E(e^{t\varepsilon})) = \sum_{n=1}^{\infty} \kappa_n(\varepsilon) \frac{t^n}{n!} = \mu_\varepsilon t + \frac{\sigma_\varepsilon^2 t^2}{2} \tag{24}$$

Apparently,  $\kappa_3(\varepsilon) = \kappa_4(\varepsilon) = \dots = 0$  for  $n \geq 3$ . Equation (24) also implies that, except for first and second cumulants, the other higher cumulants of normally distributed variables are zero.

For  $n \geq 3$ ,  $\kappa_n(Y) = g_Y^{(n)}(0) = \beta^n \kappa_n(X) + \kappa_n(\varepsilon)$ , and  $\kappa_n(\varepsilon) = 0$ . Therefore, when error term has normal distribution,  $\kappa_n(Y) = g_Y^{(n)}(0) = \beta^n \kappa_n(X)$ , and

$$\rho_{XY}^n = \beta^n \frac{(\sigma_X)^n}{(\sigma_Y)^n} = \frac{\kappa_n(Y) (\sigma_X)^n}{\kappa_n(X) (\sigma_Y)^n} = \frac{\kappa_n(Y) / (\sigma_Y)^n}{\kappa_n(X) / (\sigma_X)^n} = \frac{\gamma_n(Y)}{\gamma_n(X)} \quad (25)$$

Since  $|\rho_{XY}| \leq 1$ ,  $|\rho_{XY}|^n = \left| \frac{\gamma_n(Y)}{\gamma_n(X)} \right| \leq 1$ , it is inferred that “the  $n$ th standardized cumulant of the response, in absolute value, is always smaller than that of the explanatory variable (Dodge and Rousson, 2001)” with normally distributed error term and non-normally distributed variables.

$\gamma_3(X)$  denotes the skewness of  $X$ , defined as:

$$\gamma_3(X) = \frac{\kappa_3(X)}{\sigma_X^3} = \frac{E[(X - \mu_X)^3]}{\sigma_X^3}. \quad (26)$$

When  $n = 3$ ,  $\kappa_3(Y) = \beta^3 \kappa_3(X)$ . Thus,

$$(\rho_{XY})^3 = \beta^3 \frac{\sigma_X^3}{\sigma_Y^3} = \frac{\kappa_3(Y)}{\kappa_3(X)} \cdot \frac{\sigma_X^3}{\sigma_Y^3} = \frac{\kappa_3(Y)}{\kappa_3(X)} \cdot \frac{\sigma_Y^3}{\sigma_X^3} = \frac{\gamma_3(Y)}{\gamma_3(X)} \quad (27)$$

Since skewness provides a measure of symmetry, Equation (27) infers that the response variable tends to be more symmetric than the explanatory variables. The 4<sup>th</sup> standardized cumulant is known as the kurtosis of the distribution where

$$\gamma_4(\mathbf{X}) = \frac{\kappa_4(\mathbf{X})}{[\kappa_2(\mathbf{X})]^2} = \frac{\mu_4(\mathbf{X})}{\sigma_X^4} - 3 \quad (28)$$

Kurtosis is a measure of departure from normality. Kurtosis is zero for a normally distributed random variable. Random variables that have a positive kurtosis are called superGaussian, and those with negative kurtosis are called subGaussian. A superGaussian distribution has a higher probability of the values near the mean and the values near the extreme than the Gaussian one. On the other hand, a subGaussian distribution, e.g., uniform distribution, has a lower, wider peak around the mean and thinner tails. If  $n = 4$  is plugged into equation (25), then the result implies the response  $Y$  is closer to a normal distribution than the controlled variable  $X$ . Equation (25) is feasible to determine the distribution shape only when relevant variables have non-normal distribution, since  $\gamma_3 = \gamma_4 = 0$  for normally distributed variables.

The higher-order covariances and higher-order correlations are defined

$$\begin{aligned} \text{cov}_{jk}(X, Y) &= E\left[(X - E[X])^j (Y - E[Y])^k\right], \\ \rho_{jk}(X, Y) &= \frac{\text{cov}_{jk}(X, Y)}{(\sigma_X)^j (\sigma_Y)^k}. \end{aligned} \quad (29)$$

Since  $X$  and  $\varepsilon$  are supposed to be mutually independent, then for any two functions,  $g$  and  $h$ , we have (Hyvärinen, Karhunen and Oja, 2001):

$$E(g(X) \cdot h(\varepsilon)) = E(g(X))E(h(\varepsilon)) \quad (30)$$

Besides, when  $\text{cov}(X, \varepsilon) \neq 0$ , it infers that the relation between  $X$  and  $Y$  may come from their mutual association with another variable. Hence, when  $X$  and  $\varepsilon$  are independent, it implies that no unobserved confounding variable exists in this system (Causal Sufficiency) (Shimizu, Hyvärinen, Hoyer and Kano, 2006). We have for  $n \geq 2$  (as long as  $\text{cov}_{n0}(X, Y) \neq 0$ )

$$\begin{aligned}
\text{cov}_{n-1,1}(X, Y) &= E\left[(X - E(X))^{n-1} \cdot (Y - E(Y))\right] \\
&= E\left\{(X - E(X))^{n-1} \cdot (\alpha + \beta X + \varepsilon - E[\alpha + \beta X + \varepsilon])\right\} \\
&= E\left\{(X - E(X))^{n-1} \cdot (\beta[X - E(X)] + [\varepsilon - E(\varepsilon)])\right\} \\
&= \beta E\left[(X - E(X))^n\right] + E\left[(X - E(X))^{n-1}\right] \cdot E[\varepsilon - E(\varepsilon)] \\
&= \beta E\left[(X - E(X))^n\right] = \beta \text{cov}_{n0}(X, Y)
\end{aligned} \tag{31}$$

$$\frac{\rho_{n-1,1}(X, Y)}{\rho_{n0}(X, Y)} = \frac{\text{cov}_{n-1,1}(X, Y) / (\sigma_X)^{n-1} (\sigma_Y)}{\rho_{n0}(X, Y)} = \frac{\beta \text{cov}_{n0}(X, Y) / (\sigma_X)^{n-1} (\sigma_Y)}{\text{cov}_{n0}(X, Y) / (\sigma_X)^n} = \beta \frac{\sigma_X}{\sigma_Y} = \rho_{XY} \tag{32}$$

For  $n=3$  in (32) generates another asymmetric formula for  $\rho_{XY}$ . That is:

$$\frac{\rho_{21}(X, Y)}{\rho_{30}(X, Y)} = \frac{\rho_{21}(X, Y)}{\gamma_3(X)} = \rho_{XY} \tag{33}$$

The above equation is satisfied if and only if variable  $X$  is asymmetrically distributed so that  $\gamma_3(X) \neq 0$ . Moreover,

$$\begin{aligned}
\text{cov}_{12}(X, Y) &= E\left[(X - E(X)) \cdot (Y - E(Y))^2\right] = E\left\{(X - E(X)) \cdot (\alpha + \beta X + \varepsilon - E[\alpha + \beta X + \varepsilon])^2\right\} \\
&= E\left\{(X - E(X)) \cdot (\beta[X - E(X)] + [\varepsilon - E(\varepsilon)])^2\right\} \\
&= E\left\{(X - E(X)) \left[\beta^2(X - E(X))^2 + 2\beta(X - E(X)) \cdot (\varepsilon - E[\varepsilon]) + [\varepsilon - E[\varepsilon]]^2\right]\right\} \\
&= \beta^2 E\left[(X - E(X))^3\right] + 2\beta E\left[(X - E(X))^2\right] \cdot [E(\varepsilon) - E(\varepsilon)] + E[X - E(X)] \cdot E[\varepsilon - E[\varepsilon]]^2 \\
&= \beta^2 E\left[(X - E(X))^3\right] = \beta^2 \kappa_3(X)
\end{aligned} \tag{34}$$

In addition, with  $n=3$  in (31), we have

$$\text{cov}_{21}(X, Y) = \beta \text{cov}_{30}(X, Y) = \beta \kappa_3(X)$$

Thus,

$$\frac{\rho_{12}(X, Y)}{\rho_{21}(X, Y)} = \frac{\text{cov}_{12}(X, Y) / (\sigma_X)(\sigma_Y)^2}{\text{cov}_{21}(X, Y) / (\sigma_X)^2(\sigma_Y)} = \frac{\text{cov}_{12}(X, Y) \cdot \sigma_X}{\text{cov}_{21}(X, Y) \cdot \sigma_Y} = \frac{\beta^2 \kappa_3(X) \cdot \sigma_X}{\beta \kappa_3(X) \cdot \sigma_Y} = \beta \cdot \frac{\sigma_X}{\sigma_Y} = \rho_{XY} \quad (35)$$

By multiplying (33) and (35), the expression for the square of the correlation coefficient is obtained

$$(\rho_{XY})^2 = \frac{\rho_{12}(X, Y)}{\gamma_3(X)} \quad (36)$$

Then dividing (27) by (36) yields (as long as  $\rho_{XY} \neq 0$ )

$$\rho_{XY} = \frac{\gamma_3(Y) / \gamma_3(X)}{\rho_{12}(X, Y) / \gamma_3(X)} = \frac{\gamma_3(Y)}{\rho_{12}(X, Y)} \quad (37)$$

In summary, the correlation coefficient can be expressed by:

$$\rho_{XY} = \frac{\rho_{21}(X, Y)}{\gamma_3(X)} = \frac{\rho_{12}(X, Y)}{\rho_{21}(X, Y)} = \frac{\gamma_3(Y)}{\rho_{12}(X, Y)} \quad (38)$$

if  $\kappa_3(\varepsilon) = 0$ ,  $\rho_{XY} \neq 0$  and  $\kappa_3(X) \neq 0$

Since the square of a correlation coefficient is less than or equal to one, we have the following relation between  $X$  and  $Y$  based on formula (38) when the real model is

(14):

$$[\gamma_3(Y)]^2 \leq [\rho_{12}(X, Y)]^2 \leq [\rho_{21}(X, Y)]^2 \leq [\gamma_3(X)]^2, \quad (39)$$

In contrast, if the real condition is that  $Y$  is the cause of  $X$ , then the model could be formulated instead as:

$$X = \delta + \eta Y + \varepsilon, \quad (40)$$

where  $Y$  independent of  $\varepsilon$  and  $\varepsilon$  is normally distributed.

For  $n \geq 2$ , we therefore have

$$\begin{aligned} \text{cov}_{1,n-1}(X, Y) &= E[(X - E[X]) \cdot (Y - E[Y])^{n-1}] \\ &= E\{(\delta + \eta Y + \varepsilon - E[\delta + \eta Y + \varepsilon]) \cdot (Y - E[Y])^{n-1}\} \\ &= E\{\eta[Y - E(Y)] + [\varepsilon - E(\varepsilon)] \cdot (Y - E[Y])^{n-1}\} \\ &= \eta E[(Y - E[Y])^n] + E[(Y - E[Y])^{n-1}] \cdot E[\varepsilon - E(\varepsilon)] \\ &= \eta E[(Y - E[Y])^n] = \eta \text{cov}_{0n}(X, Y) \end{aligned} \quad (41)$$

$$\frac{\rho_{1,n-1}(X, Y)}{\rho_{0n}(X, Y)} = \frac{\text{cov}_{1,n-1}(X, Y) / (\sigma_X)(\sigma_Y)^{n-1}}{\rho_{0n}(X, Y)} = \frac{\eta \text{cov}_{0n}(X, Y) / (\sigma_X)(\sigma_Y)^{n-1}}{\text{cov}_{0n}(X, Y) / (\sigma_Y)^n} = \eta \frac{\sigma_Y}{\sigma_X} = \rho_{XY} \quad (42)$$

Taking  $n=3$  in (42) yields another asymmetric formula for  $\rho_{XY}$  when  $Y$  is treated as controlled variable and thus

$$\frac{\rho_{12}(X, Y)}{\rho_{03}(X, Y)} = \frac{\rho_{12}(X, Y)}{\gamma_3(Y)} = \rho_{XY} \quad (43)$$

Here, the explanatory variable  $Y$  is assumed to be asymmetrically distributed so that

$\gamma_3(Y) \neq 0$ . Additionally,

$$\begin{aligned} \text{cov}_{21}(X, Y) &= E[(X - E[X])^2 \cdot (Y - E[Y])] \\ &= E\{(\delta + \eta Y + \varepsilon - E[\delta + \eta Y + \varepsilon])^2 \cdot (Y - E[Y])\} \\ &= E\{(Y - E[Y])[\eta^2(Y - E[Y])^2 + 2\eta(Y - E[Y]) \cdot (\varepsilon - E[\varepsilon]) + [\varepsilon - E[\varepsilon]]^2]\} \\ &= \eta^2 E[(Y - E[Y])^3] = \eta^2 \kappa_3(Y) \end{aligned} \quad (44)$$

With  $n=3$  in (41), we have

$$\text{cov}_{12}(X, Y) = \eta \text{cov}_{03}(X, Y) = \eta \kappa_3(Y)$$

As a result,

$$\frac{\rho_{21}(X, Y)}{\rho_{12}(X, Y)} = \frac{\text{cov}_{21}(X, Y) / (\sigma_X)^2 (\sigma_Y)}{\text{cov}_{12}(X, Y) / (\sigma_X) (\sigma_Y)^2} = \frac{\text{cov}_{21}(X, Y)}{\text{cov}_{12}(X, Y)} \cdot \frac{\sigma_Y}{\sigma_X} = \frac{\eta^2 \kappa_3(Y)}{\eta \kappa_3(Y)} \cdot \frac{\sigma_Y}{\sigma_X} = \eta \cdot \frac{\sigma_Y}{\sigma_X} = \rho_{XY} \quad (45)$$

By multiplying (43) and (45), the expression for the square of the correlation coefficient is obtained

$$(\rho_{XY})^2 = \frac{\rho_{21}(X, Y)}{\gamma_3(Y)} \quad (46)$$

Moreover, the cumulants of  $X$  are

$$\kappa_n(X) = g_X^{(n)}(0) = \eta^n \kappa_n(Y) + \kappa_n(\varepsilon) = \eta^n \kappa_n(Y)$$

and

$$\frac{\gamma_3(X)}{\gamma_3(Y)} = \frac{\frac{\kappa_3(X)}{\sigma_X^3}}{\frac{\kappa_3(Y)}{\sigma_Y^3}} = \frac{\kappa_3(X)}{\kappa_3(Y)} \cdot \frac{\sigma_Y^3}{\sigma_X^3} = \eta^3 \frac{\kappa_3(Y)}{\kappa_3(Y)} \cdot \frac{\sigma_Y^3}{\sigma_X^3} = \eta^3 \cdot \frac{\sigma_Y^3}{\sigma_X^3} = \rho_{XY}^3 \quad (47)$$

Then dividing (47) by (46) yields

$$\rho_{XY} = \frac{\gamma_3(X) / \gamma_3(Y)}{\rho_{21}(X, Y) / \gamma_3(Y)} = \frac{\gamma_3(X)}{\rho_{21}(X, Y)} \quad (48)$$

In summary, when the model is  $X = \delta + \eta Y + \varepsilon$ , the correlation coefficient can be expressed by

$$\rho_{XY} = \frac{\rho_{12}(X, Y)}{\gamma_3(Y)} = \frac{\rho_{21}(X, Y)}{\rho_{12}(X, Y)} = \frac{\gamma_3(X)}{\rho_{21}(X, Y)} \quad (49)$$



Because the square of a correlation coefficient is less than or equal to one, equation (49) implies:

$$[\gamma_3(Y)]^2 \geq [\rho_{12}(X, Y)]^2 \geq [\rho_{21}(X, Y)]^2 \geq [\gamma_3(X)]^2. \quad (50)$$

The section demonstrates that we can identify the causal ordering from  $X$  to  $Y$  or its opposite by using skewness and higher-order correlations when the distributions of variables are non-normal. Independence between explanatory and disturbance variables and non-normality of variables are key assumptions in the setting. Besides, it is evident that normality of variables restricts such applications of higher-order cumulants, and thus the characteristics of non-normal distribution can be helpful to evaluate the causal ordering than the case of normally distributed variables (Dodge and Rousson, 2001).

### **Independent Component Analysis<sup>4</sup>**

The above section illustrates how to identify  $X \rightarrow Y$  over the reverse model  $Y \rightarrow X$  when we are only concerned with the connection between two variables. However, the above method works, when only considering the association between two non-normal series. If we attempt to know the causal relation among three variables or more, that is not enough. Shimizu et al. (2006) developed Linear Independent Non-Gaussian Model (LiNGAM) to do more causal search based on the assumption of independently non-Gaussian distributed disturbances by applying higher-order moment structures. LiNGAM works even when the dimension of observed non-Gaussian variables is large. LiNGAM algorithm is processed by first doing independent

---

<sup>4</sup> Parts of this section are summarized from Chapter 2 and Chapter 8 from Cover and Thomas (2006). Particularly, the statements of theorems and definitions of entropy come directly from that text.

component analysis (ICA) estimation and then discovering the correct ordering of variables.

### Basic Concept of Independent Component Analysis

Based on the theory of signal processing, the Central Limit Theorem indicates that any mixture of independent source signals usually has a distribution which is closer to normal distribution than any of the constitute source signals, even if the source signals have quite different patterns of distribution (Stone, 2004).

Assume that we observe  $n$  linear, invertible mixtures,  $x_1, \dots, x_n$ , of  $n$  independent signals,  $s_1, \dots, s_n$ .<sup>5</sup>

$$x_i = a_{i1}s_1 + a_{i2}s_2 + \dots + a_{in}s_n, \text{ for all } i. \quad (51)$$

Using the vector-matrix notation,  $X = (x_1, \dots, x_n)$  as well as  $s = (s_1, \dots, s_n)$ , the linear mixing model shown in equation (51) could be written as

$$X = As \quad (52)$$

where  $X$  as well as  $s$  are both column vectors, and  $A$  is called a “mixing matrix.” The components of  $s$  cannot be directly observed and are supposed to be mutually independent with non-Gaussian distribution. The objective of ICA is finding the “demixing matrix”  $W$  such that  $W$  maximizes the non-Gaussianity and mutual independence of the components of  $\tilde{s}$  where  $\tilde{s} = \tilde{W}X$  and  $\tilde{W} = A^{-1}$  (Hyvärinen, Karhunen and Oja, 2001; Shimizu, Hyvärinen, Hoyer and Kano, 2006; Lacerda, Spirtes, Ramsey and Hoyer, .2008).

---

<sup>5</sup> The original assumption of ICA model is that the number of observed variables must be larger or equal to the number of independent signals.

The concepts of entropy are widely applied to maximize the non-Gaussianity of signals in ICA model. The relevant terminologies will be initially introduced and then are connected with the analysis of ICA.

### Introduction of Entropy<sup>6</sup>

As Cover and Thomas indicate (2006), the entropy is a measure of the amount of information needed on the average to describe a random variable. The entropy  $H(X)$  of a discrete random variable  $X$  with  $X \in A$  and probability mass function  $p(x) = \Pr\{X = x\}$ , is defined by

$$H(X) = -\sum_{x \in A} p(x) \log p(x) \quad (53)$$

Here, the logarithm base is the number 2 and the entropy is measured in bits. Noticeably, the entropy is a function of the distribution of  $X$  instead of the actual values of  $X$ .

The expected value of a random variable  $g(X)$  is defined as

$$E[g(X)] = \sum_{x \in A} p(x) g(x) \quad (54)$$

where  $p(x)$  denotes the probability density function of  $X$ . Suppose  $g(x) = \log \frac{1}{p(x)}$ , and

then

$$\begin{aligned} H(X) &= -\sum_{x \in A} p(x) \log p(x) = \sum_{x \in A} p(x) [\log 1 - \log p(x)] \\ &= \sum_{x \in A} p(x) \left[ \log \frac{1}{p(x)} \right] \equiv E \left[ \log \frac{1}{p(X)} \right] \end{aligned} \quad (55)$$

That is the entropy of  $X$  can be interpreted as the expected value of  $\log \frac{1}{p(X)}$ .

---

<sup>6</sup> More properties of entropy and negentropy are illustrated in Appendix A.

$H(X)$  equals to 0 when  $p(X) = 0$  or 1. That means when  $p(X) = 0$  or 1,  $X$  must be a specific value or not so there is no information required to know the value of  $X$ .

### Entropy of Continuous Variables

If the concept of entropy for discrete random variables is generalized to continuous random variables case, it is called differential entropy. The differential entropy  $H(x)$  of a continuous random variable  $X$  with probability density function  $f(x)$  is defined as

$$H(X) = - \int_S f(x) \log f(x) dx \quad (56)$$

where  $S$  is the support set of the random variable.

The relative entropy  $D(f\|g)$  between two density functions  $f$  and  $g$  is defined by:

$$D(f\|g) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (57)$$

The relative entropy is a measure of the inefficiency when the true distribution,  $f(x)$ , is assumed to be  $g(x)$  incorrectly. For example, if we knew the true distribution  $f(x)$ , we could build an exact state with average needed information  $H(f(x))$ . However, if we used the code constructed on the basis of distribution  $g(x)$ , we would need

$H(f(x)) + D(f\|g)$  bits to describe  $x$ . Also, since the logarithm function,  $y = \log(x)$ , is concave, thus  $E[\log(x)] \leq \log[E(x)]$ . Consequently,

$$\begin{aligned} -D(f\|g) &= - \int f(x) \log \left[ \frac{f(x)}{g(x)} \right] dx \\ &= \int f(x) \log \left[ \frac{g(x)}{f(x)} \right] dx \leq \log \int f(x) \left[ \frac{g(x)}{f(x)} \right] dx \\ &= \log \int g(x) dx = \log 1 = 0 \end{aligned} \quad (58)$$

Intuitively,  $D(f\|g) \geq 0$ .

The joint differential entropy of a set  $X = (x_1, x_2, \dots, x_n)$  of random variables with density  $f(x_1, x_2, \dots, x_n)$  is defined as (Cover and Thomas, 2006)

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &= -\int f(x_1, x_2, \dots, x_n) \log f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_n \\ &= -\int f(X) \log f(X) dX = -E[\log f(X_1, X_2, \dots, X_n)] \end{aligned} \quad (59)$$

### Negentropy

Negentropy  $J$  is defined as

$$J(X) = J(x_1, \dots, x_n) = H(X_{gauss}) - H(X) \quad (60)$$

where  $X_{gauss}$  is a random vector of multivariate Gaussian distribution (multivariate normal distribution) with mean vector  $\mu$  and covariance matrix  $\Sigma$ .

Higher-order cumulants is a common tool used to approximate the negentropy because of the difficulty of estimating the distribution of a random variable  $x$ .

### *Hermite Polynomials and Gram-Charlier Series Expansion*

The Hermite's differential equation is given by

$$\frac{d^2 z}{dx^2} - x \frac{dz}{dx} + iz = 0 \quad (61)$$

The corresponding solutions of  $z$  can be

$$H_i(x) = (-1)^i e^{\frac{x^2}{2}} \frac{d^i}{dx^i} e^{-\frac{x^2}{2}} \quad (62)$$

where  $H_i$  is called  $i$ th-degree (probabilist's) Hermite polynomials and the order  $i$  is a nonnegative integer. If  $x$  has standardized Gaussian distribution with probability density

$\omega(x)$  such that  $\omega(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ , therefore

$$\begin{aligned} H_i(x) &= (-1)^i e^{\frac{x^2}{2}} \left( \frac{d^i}{dx^i} e^{-\frac{x^2}{2}} \right) = (-1)^i \left( \frac{d^i}{dx^i} \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) \right) / \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \right) \\ &= (-1)^i \left( \frac{\partial^i \omega(x)}{\partial x^i} / \omega(x) \right) = (-1)^i \frac{\omega^{(i)}(x)}{\omega(x)} \end{aligned} \quad (63)$$

where  $\omega^{(i)}(x)$  denote the  $i^{\text{th}}$  derivative of  $\omega(x)$ . Thus, equation (62) can be rewritten as

$$\omega^{(i)}(x) = \frac{\partial^i \omega(x)}{\partial x^i} = (-1)^i H_i(x) \omega(x) \quad (64)$$

The basic thought of Gram-Charlier Series Expansion is that the real probability density of  $x$ ,  $f(x)$ , is close to  $\omega(x)$ . Hence,  $f(x)$  can be represented by a series expansion of standardized Gaussian density function and its derivatives

$$f(x) = b_0 \omega(x) + b_1 \omega^{(1)}(x) + b_2 \omega^{(2)}(x) + \dots \quad (65)$$

For simplicity, we assume  $x$  has zero mean and unit variance. Plugging equation (64) into equation (65) yields Gram-Charlier Series Expansion such that

$$\begin{aligned} f(x) &= b_0 H_0(x) \omega(x) - b_1 H_1(x) \omega(x) + b_2 H_2(x) \omega(x) + \dots \\ &= \omega(x) [b_0 H_0(x) - b_1 H_1(x) + b_2 H_2(x) + \dots] \\ &= \omega(x) \left( \sum_{j=0}^{\infty} (-1)^j b_j H_j(x) \right) \end{aligned} \quad (66)$$

Multiply  $H_i(x)$  on both sides and integrate respect to  $x$  of equation (66). The equation becomes

$$\begin{aligned}\int f(x)H_i(x)dx &= \int \omega(x)H_i(x)\left(\sum_{j=0}^{\infty}(-1)^j b_j H_j(x)\right)dx \\ &= \sum_{j=0}^{\infty} \left[(-1)^j b_j \int \omega(x)H_i(x)H_j(x)dx\right]\end{aligned}\quad (67)$$

In addition, the orthogonal property of the Hermite polynomials is defined as

$$\int \omega(x)H_i(x)H_j(x)dx = \begin{cases} j! & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases}\quad (68)$$

When this property is applied to equation (67), the coefficients  $b_j$  can be computed as

$$b_j = \frac{(-1)^j}{j!} \int f(x)H_j(x)dx\quad (69)$$

Therefore, the first five coefficients are

$$\begin{aligned}b_0 &= \int f(x)H_0(x)dx = \int f(x)dx = 1, \\ b_1 &= -\int f(x)H_1(x)dx = \int f(x)x dx = 0, \\ b_2 &= \frac{1}{2} \int f(x)H_2(x)dx = \frac{1}{2} \int f(x)(x^2 - 1)dx = \frac{1}{2} \left[ \int f(x)x^2 dx - \int f(x)dx \right] \\ &= \frac{1}{2}(1 - 1) = 0, \\ b_3 &= -\frac{1}{3!} \int f(x)H_3(x)dx = -\frac{1}{3!} \int f(x)(x^3 - 3x)dx \\ &= -\frac{1}{3!} \left[ \int f(x)x^3 dx - 3 \int f(x)x dx \right] = -\frac{1}{3!} (\gamma_3(x) - 0) = -\frac{1}{3!} \kappa_3(x),\end{aligned}$$

$$\begin{aligned}
b_4 &= \frac{1}{4!} \int f(x)H_4(x)dx = \frac{1}{4!} \int f(x)(x^4 - 6x^2 + 3)dx \\
&= \frac{1}{4!} \left[ \left( \int f(x)x^4 dx \right) - 6 \left( \int f(x)x^2 dx \right) + 3 \left( \int f(x)dx \right) \right] = \frac{1}{4!} [\mu_4(x) - 6 + 3] \\
&= \frac{1}{4!} [\mu_4(x) - 3] = \frac{1}{4!} \kappa_4(x)
\end{aligned} \tag{70}$$

because  $x$  is standardized to have zero mean and unit variance.

Plug these coefficients into equation (66), and then  $f(x)$  can be estimated as

$$\begin{aligned}
f(x) &\approx \hat{f}(x) = b_0 H_0(x)\omega(x) - b_1 H_1(x)\omega(x) + \dots + b_4 H_4(x)\omega(x) \\
&= \omega(x) [b_0 H_0(x) - b_1 H_1(x) + \dots + b_4 H_4(x)] \\
&= \omega(x) \left( 1 \cdot H_0(x) - 0 + 0 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \\
&= \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right)
\end{aligned} \tag{71}$$

Obviously, the part of the departure from normality of  $f(x)$  is given by the third- and fourth-order cumulants in equation (71). Since the true distribution of  $x$  is presumed to be very near standardized normal distribution, the part of cumulants should be extremely small and the following approximation can be applied in the later analysis

$$\ln(1+z) \approx z - \frac{z^2}{2} \tag{72}$$

The formula of  $\hat{f}(x)$  shown in equation (71) is plugged into the definition of entropy in equation (56)

$$\begin{aligned}
H(x) &\approx - \int \hat{f}(x) \log \hat{f}(x) dx = - \int \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \\
&\ln \left[ \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \right] dx
\end{aligned}$$



$$\begin{aligned}
&= -\int \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \log \omega(x) dx \\
&- \int \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \left[ \ln \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \right. \\
&\quad \left. \overbrace{\left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right)}^{\zeta} \right] \ln \omega(x) dx \tag{73} \\
&- \int \omega(x) \left( 1 + \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right) \left[ \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right. \\
&\quad \left. - \frac{1}{2} \left( \frac{1}{3!} \kappa_3(x) H_3(x) + \frac{1}{4!} \kappa_4(x) H_4(x) \right)^2 \right] dx
\end{aligned}$$

Because of the assumption that  $f(x)$  is quite close to standardized normal distribution,

the values of  $\kappa_3(x)$  and  $\kappa_4(x)$  tend to be zero, and thus part C can be simplified to

$-\int \omega(x) \log \omega(x) dx = H(x_{gauss})$ . Besides, due to the same reason, “a third-order

monomial of  $\kappa_3(x)$  and  $\kappa_4(x)$  is infinitely smaller than terms involving only second-

order monomials (Hyvärinen, Karhunen and Oja, 2001).” After applying the above

condition and the orthogonal property in (68), the remaining part becomes

$$-\frac{(\kappa_3(x))^2}{2 \times 3!} - \frac{(\kappa_4(x))^2}{2 \times 4!} \text{ because } \int \omega(x) \kappa_i(x) H_i(x) dx = \kappa_i(x) \int \omega(x) H_i(x) H_0(x) dx = 0 \text{ for}$$

$i=3$  or  $4$ . Thus, (73) can be rewritten as

$$\begin{aligned}
H(x) &\approx -\int \omega(x) \log \omega(x) dx - \frac{(\kappa_3(x))^2}{2 \times 3!} - \frac{(\kappa_4(x))^2}{2 \times 4!} \\
&= H(x_{gauss}) - \frac{(\kappa_3(x))^2}{2 \times 3!} - \frac{(\kappa_4(x))^2}{2 \times 4!} \tag{74}
\end{aligned}$$

Therefore, the negentropy of standardized variable  $x$  is approximated by third- and fourth-order cumulants as (Choi, Grandhi and Canfield, 2006; Hyvärinen, Karhunen and Oja, 2001):

$$\begin{aligned} J(x) &= H(x_{gauss}) - H(x) \approx \frac{(\kappa_3(x))^2}{2 \times 3!} + \frac{(\kappa_4(x))^2}{2 \times 4!} \\ &= \frac{1}{12} (E(x^3))^2 + \frac{1}{48} (\gamma_4(x))^2 \end{aligned} \quad (75)$$

The objective is maximization of negentropy.

### Fast ICA Algorithm<sup>7</sup>

#### Whitening

In Fast ICA, the first step of preprocessing the observed data is to center  $X$  which means that the means of  $X$  are subtracted to be zero. Thus, the mean vector of  $s$  is calculated by  $\tilde{W}m$  where  $m$  is a mean vector of  $X$  before being centered. The next step is whitening which is that  $X$  is linearly transformed to a new vector  $\tilde{X}$  so the component of  $\tilde{X}$  are uncorrelated and their variances are equal to one,  $E(\tilde{X}\tilde{X}^T) = I$  (Hyvärinen, 2001).

Suppose  $\Sigma$  denotes the covariance matrix of the centered data  $X$ . Because  $\Sigma$  is symmetric, then we acquire a decomposition  $\Sigma = E(XX^T) = FDF^T$ , where  $F$  is an orthogonal matrix and the column vectors of  $F$  form an orthonormal basis for each eigenspace  $E_\lambda = \ker(\Sigma - \lambda I)$  and  $D$  is a diagonal matrix with eigenvalues corresponding

---

<sup>7</sup> Most texts of this section are summarized from Hyvärinen, A., Karhunen, J., and Oja E. (2001) , *Independent Component Analysis* and relevant matrix algebra operations are demonstrated in Appendix B.

to the orthonormal basis. We set  $V = FD^{-\frac{1}{2}}F^T$  which is called whitening matrix.

Whitening is implemented by:

$$\tilde{X} = VX = FD^{-\frac{1}{2}}F^T X \quad (76)$$

Obviously,

$$\begin{aligned} E(\tilde{X}\tilde{X}^T) &= E\left(\left(FD^{-\frac{1}{2}}F^T X\right)\left(FD^{-\frac{1}{2}}F^T X\right)^T\right) \\ &= E\left(\left(FD^{-\frac{1}{2}}F^T X\right)\left(X^T FD^{-\frac{1}{2}}F^T\right)\right) \\ &= FD^{-\frac{1}{2}}F^T E(XX^T)FD^{-\frac{1}{2}}F^T \\ &= FD^{-\frac{1}{2}}(F^T F)D(F^T F)D^{-\frac{1}{2}}F^T \\ &= FD^{-\frac{1}{2}}IDID^{-\frac{1}{2}}F^T = FD^{-\frac{1}{2}}DD^{-\frac{1}{2}}F^T \\ &= FF^T = I \end{aligned} \quad (77)$$

After that, another orthogonal matrix  $Q$  is necessary to be found so that the components of  $\hat{s} = Q^T \tilde{X} = Q^T VX$  have maximal non-normality and the demixing matrix is calculated by  $\hat{W} = Q^T V$  (Hyvärinen, Karhunen and Oja, 2001; Shimizu, Hoyer, Hyvärinen and Kerminen, 2006).

### Why Unobserved Signals are Presumed to be Non-Gaussian Distributed

One of the most assumptions behind ICA model is that the unobserved signals,  $s$ , has non-Gaussian distribution. Why this assumption is needed?

Suppose two independent signals,  $s = (s_1, s_2)$  are Gaussian distributed with zero mean and unit variance. Therefore, their joint probability density is:

$$f(s_1, s_2) = \frac{1}{2\pi} \exp\left(-\frac{(s_1^2 + s_2^2)}{2}\right) = \frac{1}{2\pi} \exp\left(-\frac{\|s\|^2}{2}\right) \quad (78)$$

If the mixing matrix  $A$  is orthogonal, then  $A^T = A^{-1}$  and  $|\det(A^T)| = |\det(A)| = 1$ . Hence, after the data has been whitened, the joint density of the mixtures  $X = (x_1, x_2)$  is given by:

$$\begin{aligned} f(x_1, x_2) &= f(As) = |\det(JB(s_1, s_2))|^{-1} f_{s_1, s_2}(A^{-1}X) \\ &= \frac{1}{|\det(A^T)|} \left( \frac{1}{2\pi} \right) \exp\left(-\frac{\|A^T X\|^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{\|X\|^2}{2}\right) \end{aligned} \quad (79)$$

(79) infers that the mixtures  $x_1$  and  $x_2$  are also independently Gaussian distributed.

Apparently, the distribution of orthogonal transformation of mixtures,  $A^T X$ , is identical with the distribution of the original mixture,  $X$ . The orthogonal mixing matrix  $A$  does not change the density function of  $X$  at all. It implies the mixing matrix cannot be estimated from observed variables when these variables have multivariate Gaussian distribution.

Therefore, non-Gaussianity of signals is necessary to make ICA estimation feasible

(Hyvärinen, Karhunen and Oja, 2001).

## CHAPTER III

### CAUSAL SEARCH ON THE PRICING LEADERSHIP BETWEEN THE MANUFACTURER AND THE RETAILER

#### Introduction

Discovery of linear acyclic models from purely observational data is a significant topic of current research. In this section, I will briefly review the literature on two relevant search algorithms for discovering DAGs of real data and apply such to retail-level scanner data on carbonated soft drinks (CSDs) sales and prices to examine the firm's pricing behavior.

#### Search Algorithm for Finding Causal Structure

The procedure of inferring causal structure based on probabilistic dependence is a two-step process. Step 1 establishes the probability model through statistical inferences (e.g. the parameters of means, variances, and covariances). Step 2 deduces the probabilistic (or causal) consequence from the inferential principle (e.g. d-separation) (Hoover, 2009b). Two algorithms discussed here are PC and LiNGAM. Both PC and LiNGAM algorithms<sup>8</sup> satisfy the following assumptions that: (1) the data generating process is recursive<sup>9</sup>, so its causal structure is “one-way causation,” (2) there are no omitted variables that would affect the result of causal inference so the set of observed variables is causally sufficient, and (3) the sampled observations are identically

---

<sup>8</sup> PC and LiNGAM algorithms are implemented in Tetrad which can be downloaded from the website <http://www.phil.cmu.edu/projects/tetrad/>.

<sup>9</sup> The definition of recursive is shown in Pearl's book (Pearl, 2009).

independent distributed (Shimizu, Hyvärinen, Kano and Hoyer, 2005; Hoover, 2009a).

The substantial difference between these two algorithms is the presumption of the underlying distribution of the data. PC algorithm works reliably with the normal distribution or many sorts of symmetrical but non-normal distributions, while LiNGAM algorithm assumes the non-Gaussianity of the error terms; the more non-Gaussian the better (Shimizu, Hyvärinen, Kano and Hoyer, 2005; Glymour, e-mail, 28 September 2010).

### **PC Algorithm**

PC in Tetrad<sup>10</sup> produces a pattern which represents a class of DAGs that are statistically equivalent under a normal distribution (Glymour, e-mail, 28 September 2010). When variables are multivariate Gaussian distributed, the second-order moment structure of the variables offers completely required information to define the probability density of the data and all conditional correlation of the variables can be calculated directly from their mean and covariance matrix (Shimizu, Hyvärinen, Kano and Hoyer, 2005).<sup>11</sup> This is why the PC algorithm can compute the causal ordering among variables only by using covariance matrix.

In a graphical model, two variables are connected by a line if and only if they are not conditionally independent. Since PC algorithm generates a pattern in Tetrad given the assumption that data has multivariate normal distribution, zero partial correlation or zero conditional correlation implies conditional independence (Baba, Shibata, and Sibuya, 2004). PC algorithm starts with a “complete undirected graph (Spirtes, Glymour

---

<sup>10</sup> Tetrad is software used to do causal search and it is developed by researcher in the department of philosophy at Carnegie Mellon University.

<sup>11</sup> How to derive conditional correlation of multivariate normal distributed data is shown in Appendix C.

and Scheines, 2001)” where all vertices (variables) are connected with a headless arrow. “It then examines the independence among any pairs of variables, conditional independence on sets of one variable, then two, and so forth until the set of variable is exhausted (Hoover, 2008).” If the correlation or conditional correlation is determined to be not significantly different from zero, then the edge connection between variables is removed. When all combinations for pairs of variables are tested, the direction of edges are then considered based on the relation among triples of variables. The case of inverted fork is initially identified. It considers cases where two variables,  $X$  and  $Y$ , are unconditionally independent, but related through a third variable  $Z$ . So unconditionally  $X$  and  $Y$  are independent, however if we have a variable  $Z$  such that conditional on  $Z$ ,  $X$  and  $Y$  are dependent. Then:  $X \rightarrow Z \leftarrow Y$ , and we say  $Z$  is an unshielded collider. When all unshielded colliders have been discovered, further logical rules are applied to direct other causal connections. Suppose we have  $X \rightarrow Z - Y$ , but  $Z$  was found not to be a unshielded collider previously; hence, the relation should be a causal chain such as  $X \rightarrow Z \rightarrow Y$  (Spirtes, Glymour and Scheines, 2001; Hoover, 2008).

PC algorithm may generate indistinguishable patterns due to the same conditional correlation structure (Shimizu, Hyvärinen, Kano and Hoyer, 2005). For example, in terms of Causal Markov Condition (CMC), we cannot tell  $X \rightarrow Z \rightarrow Y$  from  $X \leftarrow Z \rightarrow Y$  because both cases have the equivalent joint probability density:

$$f(X, Y, Z) = \frac{f(X, Z)f(Y, Z)}{f(Z)}. \text{ Both cases infer that } X \text{ and } Y \text{ are independent without}$$

knowledge of  $Z$  and they are independent given knowledge of  $Z$ . Actually, many patterns

of real-world data are extremely non-normal and not symmetric, and the assumption of non-Gaussian distribution is helpful to identify more causal structures which PC may not recognize. The algorithm Linear Non-Gaussian Acyclic Models (LiNGAM) applies non-Gaussian data structure for model identification.

### **Linear Non-Gaussian Acyclic Models**<sup>12</sup>

For each vector  $X = (x_1, x_2, \dots, x_n)$ ,  $x_i$  has a causal ordering. Then we could have the following structural equation model:

$$x_i = \sum_{k(j) < k(i)} b_{ij} x_j + e_i \quad (80)$$

where  $k(i)$  denotes a causal ordering and no variable listed later causes any variable listed earlier. The disturbances  $e_i$  are mutually independent and non-Gaussian distributed with non-zero variances. Equation (80) indicates that each  $x_i$  is a linear function of its preceding variables (ancestors), plus the disturbance term, but not any function of its descendents.

Initially, each variable  $x_i$  is always preprocessed by subtracting out its sample mean, to have a zero-mean vector, and, applying the vector-matrix format, equation (80) can be written as:

$$X = BX + e \quad (81)$$

---

<sup>12</sup> Parts of this section are summarized from Shimizu, Hyvärinen, Kano and Hoyer's conference paper (2005) and Shimizu, Hoyer, Hyvärinen, and Kerminen's paper (2006). Particularly, the equations of algorithm come directly from these two texts.



where  $B$  is the coefficient matrix of the model. The arrangement of  $X$  can be ordered

depending on their causal structure. Assuming  $x_3 \rightarrow x_2 \rightarrow x_1$ , then  $X = \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix}$  and hence

$B$  can be permuted to a strictly lower triangular matrix. For instance, suppose the model is defined by

$$\begin{aligned} x_3 &= e_3 \\ x_2 &= 0.5x_3 + e_2 \\ x_1 &= 0.3x_2 + e_1 \end{aligned}$$

and then

$$\begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix} \text{ where } B = \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \quad (82)$$

Solving for  $X$  in equation (80) then we obtain

$$(I - B)X = e \Rightarrow X = (I - B)^{-1}e = Ae \quad (83)$$

where  $A = (I - B)^{-1}$ .  $X$  can be expressed as a linear function of the error terms, as long as  $(I - B)$  is nonsingular. Obviously,  $A$  is lower triangular matrix with non-zero elements on the diagonal. The independence and non-normality of disturbances are assumed in (83). Equation (83) and the above presumption of disturbances form the classical linear independent component analysis (ICA) model (Hyvärinen, Karhunen and Oja, 2001; Shimizu, Hyvärinen, Hoyer and Kano, 2006). Apparently, compared to the functional format of ICA in (52), the error terms in equation (83) are often treated as “sources” or “signals” and written  $s$ . The procedures of LiNGAM algorithm include

calculating the demixing matrix  $\hat{W} = (I - \hat{B})$  by doing ICA estimation and then discovering the correct ordering of  $X$ .

*LiNGAM Discovery Algorithm*

Following the assumption behind equation (83):

$$X = Ae \text{ and equivalently } e = (I - \tilde{B})X = \tilde{W}X \quad (84)$$

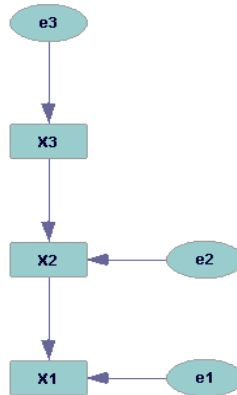
Suppose an initial  $\hat{W}$  is found by applying the FastICA algorithm introduced previously.

However, since there is no way to fix the order of independent components,  $e$  or  $s$ , the rows of initial estimated  $W$  are possibly randomly ordered. In other words, we may have the wrong correspondence between the disturbances and the observed variables.

This problem is illustrated by extending the previous example,

$$\begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix}}_{\hat{B}} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix}$$

and its corresponding causal relation is given in Figure 2.

**FIGURE 2 The Corresponding Causal Connection of Equation (82)**

The equation (82) can be written by taking error terms on the left-hand side of the equation:<sup>13</sup>

$$\begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix} = \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \right) \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0 & -0.3 & 1 \end{bmatrix}}_{\hat{W}} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} \quad (85)$$

Nevertheless, perhaps the form of initially demixing matrix  $\hat{W}$  and the corresponding order of error terms will become

$$\begin{bmatrix} e_2 \\ e_3 \\ e_1 \end{bmatrix} = \underbrace{\begin{bmatrix} -0.5 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & -0.3 & 1 \end{bmatrix}}_{\hat{W}} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} \quad (86)$$

<sup>13</sup> Equations (85) to (87) and the corresponding text are summarized the powerpoint file from The LiNGAM Project: A Longer Introduction. Available from <http://homepage.mac.com/shoheishimizu/>. Accessed July, 2010.

so therefore the correspondence between the error terms  $e_i$  and the observed variables  $x_i$  is incorrect. As the example shown in (86),  $x_3$  corresponds to  $e_2$  instead of  $e_3$  erroneously. This condition occurs is because of the “permutation indeterminacy of ICA.” In order to obtain a proper correspondence between error terms and the observed variables, a permutation matrix,  $P$ , is required to permute the rows of  $\hat{W}$ , so that there are no zeros on the main diagonal of  $\tilde{W} = P\hat{W}$  (Shimizu, 2010).

$$\begin{array}{c}
 \begin{array}{c} \begin{bmatrix} e_2 \\ e_3 \\ e_1 \end{bmatrix} = \begin{bmatrix} -0.5 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & -0.3 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} \\ \underbrace{\hspace{10em}}_{\text{if}} \end{array} \\
 \xrightarrow{\text{Permute the rows}} \\
 \begin{array}{c} \begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0 & -0.3 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} \\ \underbrace{\hspace{10em}}_{\text{if}} \end{array}
 \end{array}$$

The objective of the algorithm which searches for the row permutation matrix is to make the main diagonal elements of the demixing matrix non-zero. Therefore, this algorithm severely penalizes small absolute values of the main diagonal elements of the row-permuted demixing matrix  $\tilde{W}$ . In practice, when the number of observed variables is comparatively small (less or equal to eight)<sup>14</sup>, the algorithm is:

$$\hat{P} = \min_P \frac{1}{\sum_i \left( |P\hat{W}|_{ii} \right)} \quad (87)$$

where  $P\hat{W}$  permutes the rows of  $\hat{W}$ . Once the permutation matrix,  $\hat{P}$ , is found, we can also discover the correct demixing matrix  $\tilde{W}$  with the right correspondence between error terms and observed variables.

---

<sup>14</sup> The computation question of why eight? and not nine or seven is left unaddressed in this dissertation. Dr. Shimizu et al. indicate that it has to do with increasing computational complex algorithm at numbers higher than eight.

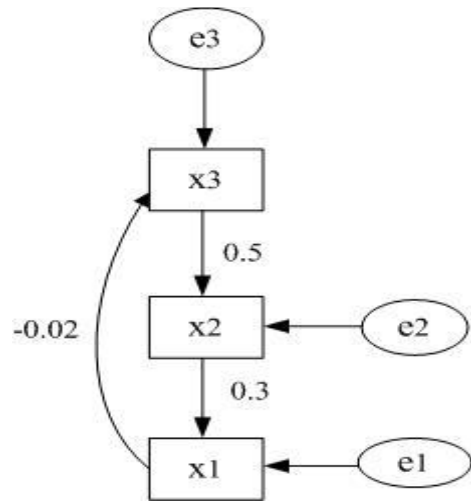
Another indeterminacy that ICA cannot solve is the scaling of the demixing matrix (or, equivalently, disturbance terms). In LiNGAM, since  $\tilde{W} = (I - \tilde{B})$  is assumed in advance, the components on the main diagonal of  $\hat{W}$ , which gives the weight of disturbance variables to the corresponding observed variables, should be fixed to one. According to Shimizu et al. (2006, pp. 2007), “each row of  $\tilde{W}$  is divided by its corresponding diagonal element,” so that the main diagonal elements of  $\tilde{W}$  are equivalent to one. Finally, the coefficient matrix  $B$  is calculated by  $\hat{B} = I - \tilde{W}$ .

The initial values of coefficients  $b_{ij}$  are estimated, but the causal ordering  $k(i)$  is yet uncertain. In other words, the first calculated coefficient matrix,  $\hat{B}$ , may not be strictly lower triangular. Besides, when applying real finite data sets on the ICA decomposition algorithm, it possibly “generates estimates which are approximately zero for those components which should be exactly zero (Shimizu, Hoyer, Hyvärinen and Kerminen, 2006).” For instance, because of the estimation error, we may get:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0.3 & 0 \\ 0 & 0 & 0.5 \\ -0.02 & 0 & 0 \end{bmatrix}}_{\hat{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} \quad (88)$$

Apparently, the corresponding causal association under this estimation is given by

Figure 3:

**FIGURE 3 The Corresponding Causal Association of Equation (88)**

This estimated result violates the structure of the DAG due to the existence of directed cycle. In order to achieve a DAG and discover the causal ordering, it is necessary to find a permutation matrix  $Q$  which permutes both rows and columns of  $\hat{B}$  simultaneously “as lower-triangular as possible” so that when the upper triangular coefficients of  $Q\hat{B}Q^T$  are set zero, the change of element’s value is the smallest. Hence, the objective is that:<sup>15</sup>

$$\hat{Q} = \min_Q \sum_{i \leq j} (Q\hat{B}Q^T)_{ij}^2 \quad (89)$$

---

<sup>15</sup> The algorithm of equations (87) and (89) are feasible only when the number of observed variables are less or equal to eight.

When the optimal permutation,  $\hat{Q}$ , is discovered, the optimal strictly lower triangular coefficient matrix  $\tilde{B}$  can be calculated by setting the upper triangular components of  $\hat{Q}\hat{B}\hat{Q}$  to zero. The estimation process of  $\tilde{B}$  is shown as (Shimizu, 2010):<sup>16</sup>

$$\begin{array}{c}
 \begin{array}{l}
 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0.3 & 0 \\ 0 & 0 & 0.5 \\ -0.02 & 0 & 0 \end{bmatrix}}_{\hat{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} \\
 \text{Simultaneously} \\
 \text{permute B} \\
 \text{as lower-} \\
 \text{triangular as} \\
 \text{possible} \\
 \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & -0.02 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix}}_{\hat{Q}\hat{B}\hat{Q}^T} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix} \\
 \text{Set upper-} \\
 \text{triangular} \\
 \text{Elements} \\
 \text{to be} \\
 \text{zeros} \\
 \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix}}_{\tilde{B}} \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_3 \\ e_2 \\ e_1 \end{bmatrix}
 \end{array}
 \end{array}$$

### *Method to Prune the Edges*

“Bentler (1983) and Mooijaart (1985) considered the generalized least squares approach in parameter estimation (Kano and Shimizu, 2003; Shimizu, Hoyer, Hyvärinen and Kerminen, 2006; Shimizu and Kano, 2008).” The developers of LiNGAM apply this methodology to evaluate the entire model fit and to direct the causal relation.

The last section shows that some coefficients of  $\hat{B}$ , implied zero given a discovered causal ordering, are set to be zero. Nevertheless, some remaining edges between variables may be weak and are “probably zero in the generating model.” Thus, a significance test of nonzero element of estimated matrix  $\tilde{B}$  (equivalently, of  $\tilde{W}$ , except the components on the main diagonal) is needed. The Wald test is used to

<sup>16</sup> Equations (88) to (89) and the corresponding text are summarized from the powerpoint file from The LiNGAM Project: A Longer Introduction. Available from <http://homepage.mac.com/shoheishimizu/>. Accessed July, 2010.

examine if some connection should be pruned away. The null and alternative hypotheses are:

$$H_0 : \tilde{w}_{ij} = 0 \quad \text{versus} \quad H_1 : \tilde{w}_{ij} \neq 0 \quad \text{for } i > j \text{ and } \tilde{w}_{ij} > 0 \quad (90)$$

identically,

$$H_0 : \tilde{b}_{ij} = 0 \quad \text{versus} \quad H_1 : \tilde{b}_{ij} \neq 0 \quad \text{for } i > j \text{ and } \tilde{b}_{ij} > 0 \quad (91)$$

so the corresponding Wald statistics is:

$$W = \frac{w_{ij}^2}{\text{avar}(w_{ij})} \quad (92)$$

where  $\text{avar}(w_{ij})$  is the estimated asymptotic variance of  $\tilde{w}_{ij}$ <sup>17</sup> and  $W$  has one degree of freedom (Bollen, 1989).

At first, the Wald statistics of each nonzero  $\tilde{w}_{ij}$ , except the main diagonal elements, is calculated. Some  $\tilde{w}_{ij}$ s are chosen when  $H_0$  of these  $\tilde{w}_{ij}$ s are not rejected. Then we set the corresponding  $\tilde{b}_{ij} = 0$  respect to  $\tilde{w}_{ij}$  with highest p-value among those chosen  $\tilde{w}_{ij}$ s. Therefore, we have a new matrix of  $\tilde{B} : \tilde{B}_{new}$ .

$X_k$  ( $k = 1, \dots, n$ ) denotes the  $k$ -th observation of a random vector  $X = (x_1, x_2, x_3)$ ,

and then the first and  $i$ -th moment structures of a random vector are defined by:

$$m_1 = \frac{1}{N} \sum_{k=1}^N X_k, \quad (93)$$

---

<sup>17</sup> The formula of  $\text{avar}(w_{ij})$  is explained in detail in the Appendix D of Shimizu, Hoyer, Hyvärinen, and Kerminen's paper (2006).



$$m_i = \text{vech} \left\{ \frac{1}{N} \sum_{k=1}^N \overbrace{[(X_k - m_1) \otimes \cdots \otimes (X_k - m_1)]}^{i \text{ times}} \right\} \quad (94)$$

where the symbol  $\otimes$  denotes the Kronecker product.

Because  $X$  is centered,  $m_1$  of the structural equation model (81) is zero and  $E(e) = 0$ . Hence, the overall fit of the model is evaluated by measuring the discrepancy between the second-order moments structure of the sample variables  $X$ ,  $m_2$ , and the model predicted second order structure  $X = (I - B)^{-1} e$ ,  $\sigma_2(\hat{\tau}_2)$ .  $m_2$  is a column vector with elements of covariance matrix of sample data itself. For example, in our modeling,

$$m_2 = (\text{var}(x_1), \text{cov}(x_1, x_2), \text{cov}(x_1, x_3), \text{var}(x_2), \text{cov}(x_2, x_3), \text{var}(x_3))$$

On the other hand, the model-based covariance matrix of centered  $X$ ,  $\Sigma$ , can be re-written as:<sup>18</sup>

$$\begin{aligned} \Sigma &= E(XX^T) = E \left[ (I - B)^{-1} e \left( (I - B)^{-1} e \right)^T \right] = E \left[ (I - B)^{-1} e e^T \left( (I - B)^{-1} \right)^T \right] \\ &= (I - B)^{-1} \text{cov}(e) \left( (I - B)^{-1} \right)^T = (I - B)^{-1} D \left( (I - B)^{-1} \right)^T \\ &= (I - B)^{-1} D^{\frac{1}{2}} D^{\frac{1}{2}} \left( (I - B)^{-1} \right)^T = \left[ D^{-\frac{1}{2}} (I - B) \right]^{-1} \left\{ \left[ D^{-\frac{1}{2}} (I - B) \right]^{-1} \right\}^T \\ &= \mathbf{Y} \mathbf{Y}^T \end{aligned} \quad (95)$$

where  $D = E(ee^T) = \text{cov}(e)$  and  $\mathbf{Y} = \left[ D^{-\frac{1}{2}} (I - B) \right]^{-1}$ . Because the disturbance terms are

assumed to be independent of each other,  $D$  is a diagonal matrix. Furthermore,

---

<sup>18</sup> Relevant matrix algebra operations are demonstrated in Appendix B.

$$\sigma_2(\hat{\tau}_2) = \left( (YY^T)_{11}, (YY^T)_{21}, (YY^T)_{31}, (YY^T)_{22}, (YY^T)_{32}, (YY^T)_{33} \right)$$

where  $\tau_2$  contains the nonzero estimated coefficients of  $B$  and the elements of  $D$ . In our case,

$$\tau_2 = (b_{21}, b_{32}, d_{11}, d_{22}, d_{33})$$

Then the null and alternative hypotheses of testing the overall model fit are:

$$H_0 : E(m_2) = \sigma_2(\tau_2) \quad \text{versus} \quad H_1 : E(m_2) \neq \sigma_2(\tau_2) \quad (96)$$

Define

$$F(\hat{\tau}_2) = \{m_2 - \sigma_2(\hat{\tau}_2)\}^T \hat{M} \{m_2 - \sigma_2(\hat{\tau}_2)\} \quad (97)$$

where

$$\begin{aligned} \hat{M} &= \hat{V}^{-1} - \hat{V}^{-1} \hat{J} (\hat{J}^T \hat{V}^{-1} \hat{J})^{-1} \hat{J}^T \hat{V}^{-1} \\ \hat{J} &= \frac{\partial \sigma_2(\hat{\tau}_2)}{\partial \hat{\tau}_2^T}. \end{aligned} \quad (98)$$

$V$  is the covariance matrix of  $m_2$ , and  $n$  is the number of the observations.<sup>19</sup> Although a test statistic  $T_1 = n \times F(\hat{\tau}_2)$  could be used to test the null hypothesis displayed in (96), we generally require large sample sizes for  $T_1$  to have an approximately  $\chi^2$  distribution, so relevant studies of LiNGAM apply the test statistic  $T_2$  from Yuan-Bentler's suggestion (Yuan and Bentler, 1997):

$$T_2 = \frac{T_1}{1 + F(\hat{\tau}_2)} \quad (99)$$

---

<sup>19</sup> The exact form of estimated  $J$  refers to the Appendix E of Shimizu, Hoyer, Hyvärinen, and Kerminen's paper (2006) and is illustrated in Appendix D of my dissertation.

“ $T_2$  has an approximately  $\chi^2$  distribution with degrees  $u - v$  of freedom where  $u$  is the number of the distinct moments and  $v$  is the number of elements of  $\tau_2$  (Shimizu, Hoyer, Hyvärinen and Kerminen, 2006).”  $T_2$  only can be applied when its degree of freedom is larger than zero.

Now consider that Model 1 employs  $\tilde{B}$  with  $r$  edges while Model 2 employs  $\tilde{B}_{new}$  with  $r - 1$  edges when estimating the model predicted second-order moment structure. Suppose  $T_2(r)$  and  $T_2(r - 1)$  are the statistic  $T_2$  for Model 1 and Model 2, respectively. The absolute value of  $T_2(r) - T_2(r - 1)$  is also asymptotically approximately  $\chi^2$  distributed. If the null hypothesis  $H_0 : T_2(r) - T_2(r - 1) = 0$  is not rejected and  $T_2(r - 1)$  does not reject the null hypothesis of overall model fit shown in (96), it means there is no significant difference in a model fit when this edge is removed. In general, a simpler model is preferred, so  $\tilde{B}_{new}$  is accepted to substitute the original  $\tilde{B}$  and the weakest nonzero  $\tilde{b}_{ij}$  is pruned out. Otherwise, the original  $\tilde{B}$  is accepted. Next, the second-weakest edge is tested until all selected  $\tilde{w}_{ij}$ s with non-significant Wald statistics are exhausted (Shimizu, Hoyer, Hyvärinen and Kerminen, 2006). Finally, the optimal  $\tilde{B}$  (or  $\tilde{W}$ ) is evaluated and, thus, we can have a well-fitting overall model.

#### *Determine the Direction of Causality*

Although statistic  $T_2$  can be used to test the fitness of predicted model, however for some simple modeling, only applying second moment structures is not sufficient. For example, suppose

$$\text{Model 1}' : y = \beta x + \varepsilon_y$$

$$\text{Model 2}' : x = \eta y + \varepsilon_x$$

where  $x$  and  $y$  are centered. Further write moments:

$$m_{ij} = \frac{1}{N} \sum_{k=1}^N x_k^i y_k^j \quad (100)$$

The first-order moments of sample data, the expected values of  $x$  and  $y$ , are not considered here because the sample data are centered, so  $E(x) = E(y) = 0$ .

Suppose Model 1' holds true where  $x$  and  $\varepsilon_y$  are independent, then the corresponding model predicted second-order moment structure is given:

$$E \begin{bmatrix} m_{20} \\ m_{11} \\ m_{02} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \\ \beta^2 & 1 \end{bmatrix} \begin{bmatrix} E(x^2) \\ E(\varepsilon_y^2) \end{bmatrix} = \sigma_2(\hat{\tau}_2) \quad (101)$$

where  $\hat{\tau}_2 = (\beta, E(x^2), E(\varepsilon_y^2))$ . Obviously, the number of the distinct sample moments is identical to the number of  $\tau_2$ . On the other hand, the second-order moments of Model 2' are the same as Model 1'. It is said that Model 1' and Model 2' are "saturated" and  $T_2$  cannot be used to evaluate which model has a better model fit. Accordingly, the conclusion is the same as we explained in Chapter II; Model 1' cannot be identified from Model 2' if only second-order moment structures are concerned.

Nevertheless, since the relevant variables and disturbance terms are assumed to be non-normally distributed, if the corresponding third- and fourth-order moments of Model

1' and Model 2' are different from each other, the moments up to fourth order can be applied to detect the causal direction.<sup>20</sup>

Following the hypotheses in (96), we extend the moment structures, used to evaluate an overall model fit, up to the fourth-order moment. Let us denote

$$m = [m_2^T, m_3^T, m_4^T]^T \text{ and } \sigma(\hat{\tau}) = [\sigma_2(\tau_2)^T, \sigma_3(\tau_3)^T, \sigma_4(\tau_4)^T]^T \quad (102)$$

where  $m$  includes components of the second- to fourth-order moments of sample data and  $\sigma(\hat{\tau})$  consists of the elements of the second- to fourth-order moments of the model considered to data.

Therefore, the null and alternative hypotheses of testing the overall model fit become:

$$H_0 : E(m) = \sigma(\tau) \quad \text{versus} \quad H_1 : E(m) \neq \sigma(\tau) \quad (103)$$

The difference between the moments structures of sample data and assumed model is examined not only by using second-order moments alone but also applying up to fourth-order moments. Thus, the corresponding test statistic  $F(\hat{\tau})$ ,  $T_1$  and  $T_2$  change into

$$F(\hat{\tau}) = \left\{ \begin{bmatrix} m_2 \\ m_3 \\ m_4 \end{bmatrix} - \begin{bmatrix} \sigma_2(\hat{\tau}_2) \\ \sigma_3(\hat{\tau}_3) \\ \sigma_4(\hat{\tau}_4) \end{bmatrix} \right\}^T \hat{M} \left\{ \begin{bmatrix} m_2 \\ m_3 \\ m_4 \end{bmatrix} - \begin{bmatrix} \sigma_2(\hat{\tau}_2) \\ \sigma_3(\hat{\tau}_3) \\ \sigma_4(\hat{\tau}_4) \end{bmatrix} \right\} \quad (104)$$

$$T_1 = n \times F(\hat{\tau})$$

$$T_2 = \frac{T_1}{1 + F(\hat{\tau})}$$

Suppose, compared to Model 2', Model 1' has a lower  $T_2$  and does not reject  $H_0$  shown in (103). Then it implies that Model 1' has better model-data consistency so it is the

---

<sup>20</sup> Related proofs of higher-order moments are shown in Appendix E.

best-fitting model. Therefore, Model 1' reflects the correct causal ordering among variables (Kano and Shimizu, 2003; Shimizu and Kano, 2008).

The above introduction does not concern the structure of time series data. However, in my application, not only “instantaneous effect” but also “lagged influences” between time series data  $x_{it}$  are considered.

### Structural Vector Autoregressive Model with LiNGAM

Following Hyvärinen, Zhang, Shimizu, and Hoyer's studies (2010), suppose that there are  $n$  related variables,  $X_t = (x_{1,t}, \dots, x_{n,t})$ , and the vector autoregression model combined with the framework of LiNGAM can be defined as:

$$X_t = B_0 X_t + B_1 X_{t-1} + \dots + B_p X_{t-p} + e_t \quad (105)$$

where  $p$  is the number of time lag used, and  $B_i = \begin{bmatrix} b_{11,i} & \dots & b_{1n,i} \\ \vdots & \ddots & \vdots \\ b_{n1,i} & \dots & b_{nn,i} \end{bmatrix}$ .  $B_0$  shows the

“instantaneous effects” and reflects the causal orderings of variables.  $B_0$  plays a role as  $B$  in equation (81) and  $B_0$  should be restricted to be a strictly lower triangular matrix.

Moreover,  $B_i$  indicates the impact from the past to the current time for  $i \geq 1$ . The

disturbances  $e_t$  have the same properties as what are defined in LiNGAM. Equation

(105) can be rewritten as:

$$(I - B_0)X_t = B_1 X_{t-1} + \dots + B_p X_{t-p} + e_t = \sum_{i=1}^p B_i X_{t-i} + e_t \quad (106)$$

This equation becomes

$$X_t = \sum_{i=1}^p (I - B_0)^{-1} B_i X_{t-i} + (I - B_0)^{-1} e_t = \sum_{i=1}^p M_i X_{t-i} + (I - B_0)^{-1} e_t \quad (107)$$

where  $M_i = (I - B_0)^{-1} B_i$ . The way to recover the causal matrices  $B_i$  is to do vector autoregressive (VAR) model estimation on (107) and then we can have the estimated autoregressive matrices  $\hat{M}_i$ . Hence, the residuals are calculated by

$$\hat{u}_t = (I - B_0)^{-1} e_t = X_t - \sum_{i=1}^p \hat{M}_i X_{t-i} \quad (108)$$

Since

$$\hat{u}_t = (I - B_0)^{-1} e_t \Rightarrow (I - B_0) \hat{u}_t = e_t \Rightarrow \hat{u}_t = B_0 \hat{u}_t + e_t \quad (109)$$

Therefore, LiNGAM estimation is performed in (109) to discover the matrix  $B_0$ . Other

$B_i$ s are calculated by

$$B_i = (I - \hat{B}_0) \hat{M}_i \text{ for } i \geq 1 \quad (110)$$

(Hyvärinen, Zhang, Shimizu, and Hoyer 2010).

### Strategic Interaction between Firms

Economic theory alone cannot tell us which strategies that firms use in the real world and therefore empirical research of a firm's behavior has received considerable attention because many researchers are interested in realizing how firms truly behave (Perloff, Karp, and Golan, 2007).

If manufacturers and retailers repeated their interaction following a specific pattern, the application of causality analysis on their pricing interaction can make their strategic

behavior clear and the DAG result can describe the firms' relation that is in equilibrium (Dominguez, 2009). Moreover, another advantage of causal analysis is that the pricing interaction between firms can be investigated without imposing any given structure a priori.

When discussing the interactions between manufacturer and retailer, a vertically-integrated system and Stackelberg leadership of bilateral-monopoly modeling are the main cases considered. A vertically-integrated system describes a situation when the manufacturer and retailer cooperate to work as an integrated firm and aims to maximize the profit of the entire channel instead of individual benefit respectively. Also, the manufacturer's price is viewed as a cost in this system and then they share the total margins in the distribution channel. Under this circumstance, the important character of this model is that both the manufacturer price and the retail price affect the sales condition regardless of the relation between  $p_m$  and  $p_r$ . Therefore, such game may

imply the graphs as  $\begin{matrix} p_r \rightarrow & p_m \\ \searrow & \swarrow \\ & q \end{matrix}$  or  $\begin{matrix} p_r \leftarrow & p_m \\ \searrow & \swarrow \\ & q \end{matrix}$  (Dominguez, 2009).

In Stackelberg leadership modeling, the Stackelberg leader is assumed to have the ability to envision how its opponents will react in response to his strategy while the follower is unable to know how his behavior affects the leader's strategic choice. Thus, the Stackelberg leader can gain a larger share of the overall channel profits and has stronger pricing power than the follower (Ingene, and Parry, 2004). Furthermore, the manufacturer's price can be manipulated by the retail price in Retailer Stackelberg game, whereas the retailer's price would be affected by the setting of the manufacturer's price



in Manufacturer Stackelberg game (Dominguez, 2009). Because the price leader chooses its own price to maximize its profit, rather than the whole channel's profit, the channel profit is less than the one under vertically-integrated system (Ingene, and Parry, 2004).

In this study, we observe the causal relation among manufacturer's selling price, retail price and sales quantity to examine if there is a vertically-integrated connection between the retailer and CSD manufacturers or if the Stackelberg leadership is controlled by the CSD manufacturer, or by the retailer (Dominguez, 2009).

## **Data and Empirical Result**

### **Database Description**

Highly disaggregate data at frequent observation intervals are properly used to figure out the structure of repeated-game strategies (Slade, 1992). The main data resource for this study is the public Dominick Database from the Kilts Center for Marketing at the University of Chicago's Booth School of Business. The scanner database contains weekly retail prices, the number of packages sold, and gross margin information for more than 3500 UPCs for over 100 stores operated by Dominick's Finer Foods (DFF) in the Chicago metropolitan area. We select the product list in the Soft Drinks category.

Since VAR-LiNGAM estimation is another important issue in this essay, continuously observed time series data are needed. However, because this dataset has missing records during weeks #254 to #261, only data from weeks #1 to #253 (09/14/89-07/20/94) is used in this study. Additionally, since there are also no records on week #24

and week #211 for most goods, total number of observations for each CSD is 251. The simulation result of Shimizu, Hyvärinen, Hoyer and Kano's paper (2006) indicates that about 80% of causal orderings are recovered when the trial number equals 250. In order to further simplify the data into time-ordered series rather than panel data structure, only sales records of store #111 are analyzed. Store #111 is located in Chicago, IL 60620. This database has no entries if arbitrary goods were not sold in a certain week. Thus, store #111 is selected because it has comparably complete sales record of the goods relative to records of other stores we examined. The advantage of this database is that the prices charged by the manufacturers can be derived through the provided gross margin measure. This characteristic is helpful for us to explore the relationship between carbonated soft drink (CSD) manufacturers and one retailer in a supply chain.

### Characteristics of Data

The top-ten best-selling CSD brands with 6-pack, 12-pack, 24-pack, and 2 liters between 1996 and 1998 are considered in this study, including Coke Classic, Pepsi-Cola, Diet Coke, etc. The triple variables examined here consist of  $(q_i, p_{r,i}, p_{m,i})$  where

Notation	Description
$q_i$	The number of packages sold of CSD $i$ .
$p_{r,i}$	Retailer's selling price of CSD $i$ .
$p_{m,i}$	Selling price of CSD $i$ charged by the manufacturer.

The manufacturers are possibly the syrup producers, such as Coca-Cola Company and PepsiCo, or distributed bottlers. Missing entries due to no sales in a given week do

occur, so only drinks with more than 240 observations are taken into account. Thus, we study 27 products. We reject the null hypothesis of non-stationarity by using Phillips-Perron test for most series. These series look stationary. Some price series which do not reject Phillips-Perron test are verified to reject the Augmented Dickey-Fuller test with constant drift at a 0.05 significance level. Stationarity of the series avoids the possibility of spurious results. In addition, all series, except for the  $p_r$  of Canada Dry 6-pack and the  $p_m$  of Canada Dry 12-pack, reject normality with Kolmogorov-Smirnov test<sup>21</sup> and about 90% of series reject the symmetry test, which help demonstrate that it is proper to apply LiNGAM to investigate the causal connection among these series. Generally, the series of the number of packages sold always has the highest kurtosis while the retail price series has the lowest kurtosis for each product. The average kurtosis of  $(q, p_r, p_m)$  are (32.84, -0.518, 2.08) respectively. Moreover, the series of packages sold still have the highest skewness but the wholesale price series have the lowest skewness rather than the retail price series. The average skewness of  $(q, p_r, p_m)$  are (4.805, -0.495, -1.347) respectively. The only information this data reveals is the series of  $q$  is obviously far from normal distribution.

### **Prune Factor**

The prune factor is essentially a heuristic parameter that plays a role very similar to that of a significance level for the PC algorithm. Generally, the prune factor is used to decide how easily weak connections are pruned away. If the prune factor is equal to 0, there is no further pruning for the matrix of estimated connection strengths,  $B$ . The larger

---

<sup>21</sup> The Kolmogorov-Smirnov test is exercised in matlab.

the prune factor is, the more edges would be pruned out. However, there is not yet a standard pruning method specialized to LiNGAM, according to the LiNGAM developers. The prune factor approach is a simplified version of bootstrapping. To Dr. Shimizu's understanding, the approach is implemented mainly for computational efficiency and it might not have very strong theoretical support. He suggests that it might be better to do bootstrapping to see if the connection between two variables is significant. This would involve fixing the variable ordering to be the estimated one by LiNGAM and doing ordinary least squares on the bootstrap samples (Shimizu, e-mail, 16 September 2010).

In contrast to the prune factor in LiNGAM, the significance level should decrease when the sample size increases in order to derive the correct result. Spirtes et al. suggest the proper significance level should be 0.1 with sample sizes between 100 and 300 (Spirtes, Glymour, and Scheines 2001). In our application, most of the sample sizes are equal to or a little bit lower than 251, so it is reasonable to use 0.05 and 0.1 to be the significance level in PC.

### **Empirical Results**

The default value of prune factor in the matlab package of LiNGAM is 1. Therefore, we use this value of prune factor when applying LiNGAM and use 0.05 to be the significance level when applying PC algorithm; then compare the resultant graphs under these two algorithms simultaneously. The results of the estimates are presented in Table 1. If there is no connection between variables, it implies that the corresponding relation strength is zero,  $b_{ij} = 0$ , in LiNGAM, while it means that the covariance of

these two variables is zero,  $\text{cov}(x_i, x_j) = 0$ , in PC. 100% causal relations can be determined in LiNGAM although the resultant relationships between  $p_r$  and  $p_m$  of 7 Up and Canada Dry are not stable. The unstable results of Canada Dry may come from the distribution of relevant series. For example, the series of Canada Dry 6-pack's retail price does not reject the null hypothesis of symmetry and normality test. Expanding the sample size possibly can bring a solution. More observations are preferred because more data provides more accurate LiNGAM estimations. The simulation outcome of LiNGAM shows that more than 95% of causal orderings are correctly recovered when the sample size is more than 500 (Shimizu, Hyvärinen, Hoyer and Kano, 2006). On the other hand, the variables are only connected by headless arrows in PC because PC generates several DAGs that are statistically equivalent based on the assumption of a normal distribution of the sample data and it does not offer the exact directed causal flow in our case. Such results demonstrate that the non-Gaussianity of data is a helpful aide in model identification compared to the presumption of normality.

Furthermore, the LiNGAM results indicate 74% of the products have a pricing pattern like Retailer Stackelberg leadership,  $p_r \rightarrow p_m$ .

**TABLE 1 Empirical Graphs of LiNGAM and PC Estimates for CSDs**

Products/Algorithm	LiNGAM (Prune Factor=1)	PC (Significance Level=0.05)
Coke Classic		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
24-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q$
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
Pepsi Cola		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $q$
12-pack	$p_r \leftarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
24-pack	$p_r \rightarrow p_m$ $\searrow q$	$p_r - p_m$ $\setminus q /$
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
Diet Coke		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$
Mountain Dew		
2 liters	$p_r \rightarrow p_m$ $q \swarrow$	$p_r - p_m$ $\setminus q /$
Sprite		
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q$
Dr Pepper		
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\setminus q /$

TABLE 1 Continued		
Products/Algorithm	LiNGAM (Prune Factor=1)	PC (Significance Level=0.05)
Diet Pepsi		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $q$ /
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
7 Up		
6-pack	$p_r \rightarrow p_m$ or $p_r \leftarrow p_m$ $q$ $q$	$p_r - p_m$ \ $q$ /
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
24-pack	$p_r \rightarrow p_m$ or $p_r \leftarrow p_m$ $q$ $q$	$p_r - p_m$ \ $q$ /
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
CF Diet Pepsi		
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
Diet 7 Up		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
2 liters	$p_r \leftarrow p_m$ $q$	$p_r - p_m$ \ $q$ /
Canada Dry		
6-pack	$p_r \leftarrow p_m$ or $p_r \rightarrow p_m$ $q$ $q$	$p_r - p_m$ \ $q$ /
12-pack	$p_r \leftarrow p_m$ $q$	$p_r - p_m$ $q$ /

TABLE 1 Continued		
Products/Algorithm	LiNGAM (Prune Factor=1)	PC (Significance Level=0.05)
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $q /$

At least, it is clear that the retailer has the ability to affect the setting of the price charged by the manufacturers of Coke, DietPepsi, and Caffeine-Free Diet Pepsi (CF Diet Pepsi) regardless of the package sizes. This finding is consistent with Kadiyali, Chintagunta, and Vilcassim's (2000) conclusion. Their study calculates the pricing power, in terms of the gross-margin, in the distribution channel. They also study Dominick's scanner data for the period 09/14/89-11/25/93 which is very similar to the time period we use. Their conclusion is that although the brand with higher market share, Tropicana, has higher estimated manufacturer channel profit share than the brand with lower market share, MinuteMaid, the retailer is evaluated to own 62.35% of total channel profit on average in the refrigerated juice category.

On the other hand, PC gives a set of undetermined connection graphs: 70.4% of these graphs describe that  $q$ ,  $p_r$ , and  $p_m$  are related to each other but the connection pattern is uncertain. Furthermore, although the connections of some products do not appear in LiNGAM, PC's outcomes show that the relationship among these variables exists inversely, such as series of 12-pack Coke Classic, or 6-pack Diet Coke. There is no exact method we can offer here to prove which algorithm offers more accurate patterns. In general, our results support that LiNGAM is a quite different algorithm with



different properties from PC (a result consistent with recent observations of Glymour, e-mail, 28 September 2010).

The flow  $p_r \rightarrow q$  is always anticipated, but we hardly see this outcome in LiNGAM's results. The relations between prices are mostly elicited while the relation between the number of packages sold and prices are not so apparent when the prune factor is supposed to be one.

Table 2 shows the estimated causal ordering from equation  $\hat{u}_t = B_0 \hat{u}_t + e_t$  when the lagged effects are considered. Empirically, all of the residuals  $\hat{u}_t$  reject the null hypothesis of symmetry and normality tests. Hence, these residuals should be non-Gaussian distributed. Schwarz Information Criterion is used to choose the optimal time lag for the best multivariate time series fit. Finally, only 10 products have lagged effects. After comparing the outcomes of Table 1 and Table 2 for the same products, generally, the resultant causal orderings for each good are quite similar. There is no apparent change in the causal interaction between retail price and manufacturer's price for each product. However, more cases of the retail price that stimulates the sales condition,  $p_r \rightarrow q$ , are drawn forth in LiNGAM.

**TABLE 2 Causal Associations of Residuals from VAR-LiNGAM and VAR-PC Estimates for CSDs**

Products/Algorithm	Lags	VAR-LiNGAM (Prune Factor=1)	VAR-PC (Significance Level=0.05)
Coke Classic			
6-pack	1	$  \begin{array}{c}  u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
24-pack	1	$  \begin{array}{c}  u_{p_r} \quad u_{p_m} \\  \quad \searrow \quad \swarrow \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
Pepsi Cola			
24-pack	1	$  \begin{array}{c}  u_{p_r} \quad u_{p_m} \\  \quad \searrow \quad \swarrow \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
Diet Coke			
6-pack	1	$  \begin{array}{c}  u_{p_r} \quad u_{p_m} \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
7 Up			
24-pack	1	$  \begin{array}{c}  u_{p_r} \leftarrow u_{p_m} \\  \quad \searrow \quad \swarrow \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
CF Diet Pepsi			
12-pack	2	$  \begin{array}{c}  u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
2 liters	1	$  \begin{array}{c}  u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
Canada Dry			
6-pack	1	$  \begin{array}{c}  u_{p_r} \leftarrow u_{p_m} \quad \text{or} \quad u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q \quad \quad \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
12-pack	1	$  \begin{array}{c}  u_{p_r} \leftarrow u_{p_m} \quad \text{or} \quad u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q \quad \quad \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $
2 liters	2	$  \begin{array}{c}  u_{p_r} \rightarrow u_{p_m} \\  \quad \quad u_q  \end{array}  $	$  \begin{array}{c}  u_{p_r} - u_{p_m} \\  \quad \quad u_q  \end{array}  $

The differences in causal structures from LiNGAM among 6-pack, 24-pack Coke and 24-pack Pepsi are perhaps indicative of impulse buying for the smaller quantity size 6-pack relative to 24-pack. The latter involves more monetary out lags than the former.

Relaxing the strength of pruning may provide more causal relation as the economics theory refers. Therefore, we also examine the causal pattern when the prune factor in LiNGAM is set to be 0.5 and significance level in PC is set to be 0.1. The relevant results are shown in Table 3. For PC algorithm's outcomes, there is no obvious change. As expected, more edges are remained and up to 81.5% of the graphs represent the mutual correlation between  $q$ ,  $p_r$ , and  $p_m$ . On the contrary, a lot of causal patterns of  $p_r \rightarrow q$ ,  $p_m \rightarrow q$ , and even  $p_m \rightarrow q \leftarrow p_r$  appear in LiNGAM's outcomes.  $p_r \rightarrow q$  is always assumed in most demand analysis and our results show the existence of such connection. Further,  $p_m \rightarrow q$  is also very common to see in Table 3. It is not easy to explain why such condition exists directly. One possible reason is the scarcity of shelf space. CSDs industry is a highly competitive so the prices charged by manufacturer to retailer can become a good tool to pursue a better shelf location to appeal shopper's attention. Moreover,  $p_m \rightarrow q \leftarrow p_r$  implies the possible cooperation between manufacturer and retailer to maximize their joint profits in a vertically-integrated system. The residual with looser pruning criterion is also examined. Primarily, the results present more relations between manufacturer's price and the number of package sold,  $p_m \rightarrow q$ , as shown in Table 3.

**TABLE 3 Empirical Graphs of LiNGAM and PC Estimates for CSDs with Lower Prune Factor and Significance Level**

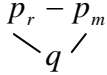
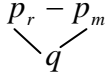
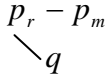
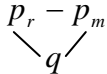
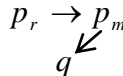
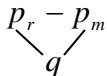
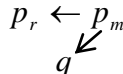
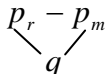
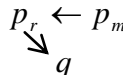
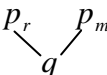
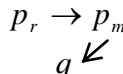
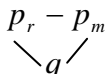
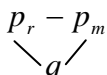
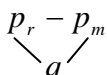
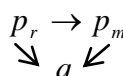
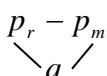
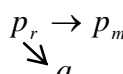
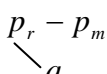
Products	LiNGAM (Prune Factor=0.5)	PC (Significance Level=0.1)
<b>Coke Classic</b>		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ 
12-pack	$p_r \quad p_m$ $q$	$p_r - p_m$ 
24-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ 
2 liters	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ 
<b>Pepsi Cola</b>		
6-pack	$p_r \rightarrow p_m$ 	$p_r - p_m$ 
12-pack	$p_r \leftarrow p_m$ 	$p_r - p_m$ 
24-pack	$p_r \leftarrow p_m$ 	$p_r - p_m$ 
2 liters	$p_r \rightarrow p_m$ 	$p_r - p_m$ 
<b>Diet Coke</b>		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ 
2 liters	$p_r \quad p_m$ $q$	$p_r - p_m$ 
<b>Mountain Dew</b>		
2 liters	$p_r \rightarrow p_m$ 	$p_r - p_m$ 
<b>Sprite</b>		
2 liters	$p_r \rightarrow p_m$ 	$p_r - p_m$ 

TABLE 3 Continued		
Products	LiNGAM (Prune Factor=0.5)	PC (Significance Level=0.1)
Dr Pepper		
2 liters	$p_r \rightarrow p_m$ $q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$
Diet Pepsi		
6-pack	$p_r \rightarrow p_m$ $q \swarrow$	$p_r - p_m$ $q \swarrow$
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\swarrow q \searrow$
2 liters	$p_r \rightarrow p_m \quad p_r \leftarrow p_m$ $q \swarrow \quad \text{or} \quad q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$
7 Up		
6-pack	$p_r \leftarrow p_m \quad \text{or} \quad p_r \rightarrow p_m$ $q \quad \quad \quad q$	$p_r - p_m$ $\swarrow q \searrow$
12-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\swarrow q \searrow$
24-pack	$p_r \rightarrow p_m \quad p_r \leftarrow p_m$ $\swarrow q \quad \quad \quad \swarrow q$	$p_r - p_m$ $\swarrow q \searrow$
2 liters	$p_r \leftarrow p_m$ $q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$
CF Diet Pepsi		
12-pack	$p_r \rightarrow p_m$ $\swarrow q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$
2 liters	$p_r \rightarrow p_m$ $\swarrow q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$
Diet 7 Up		
6-pack	$p_r \rightarrow p_m$ $q$	$p_r - p_m$ $\swarrow q \searrow$
2 liters	$p_r \leftarrow p_m$ $q \swarrow$	$p_r - p_m$ $\swarrow q \searrow$

TABLE 3 Continued		
Products	LiNGAM (Prune Factor=0.5)	PC (Significance Level=0.1)
Canada Dry		
6-pack	$p_r \leftarrow \begin{matrix} p_m \\ \swarrow \\ q \end{matrix}$ or $p_r \rightarrow \begin{matrix} p_m \\ \swarrow \\ q \end{matrix}$	$\begin{matrix} p_r - p_m \\ \swarrow \\ q \end{matrix}$
12-pack	$p_r \leftarrow \begin{matrix} p_m \\ \swarrow \\ q \end{matrix}$	$\begin{matrix} p_r - p_m \\ \swarrow \\ q \end{matrix}$
2 liters	$p_r \rightarrow \begin{matrix} p_m \\ \swarrow \\ q \end{matrix}$	$\begin{matrix} p_r - p_m \\ \swarrow \\ q \end{matrix}$

### Conclusion

The most common LiNGAM resultant graphs, for the soft drinks which we study, indicate that the retail selling price can affect the manufacturer's charge for CSDs. Retailer's pricing power may come from its market share in supermarket industry or intensive competition among Coca-Cola, PepsiCo, and Cadbury (Kadiyali, Chintagunta, and Vilcassim, 2000). However, there is no proper way to prove if the above conjecture is correct. Although there are no other studies that can verify our results directly, the related research on manufacturer-retailer channel interactions in the drinks categories may offer some supports. For example, Kadiyali, Chintagunta, and Vilcassim (2000) set up a structural model to estimate the manufacturer and retailer have how much market pricing power respectively. Their results indicate that the retailer has a higher channel profit share than manufacturer in refrigerated juice product category. Also, the retailer has stronger pricing power, as measured by markup, rather than manufacturer for each national brand of refrigerated juice. Dominick database includes retailer's gross margin but lacks for the information of manufacturer's margin. If the manufacturer's gross

margin is available, that can provide another verification way of our result. However, they also indicate that the proposed vertical Nash, manufacturer Stackelberg, and retailer Stackelberg models are all rejected so it implies the real channel interaction is more complicated than the traditionally bilateral-monopoly models. Therefore, although our LiNGAM estimation makes the causal associations among manufacturer's price, retailer's price and sold amount appear, it may not be proper to conclude that such result can represent a specific pricing game as theory proposed. At least, our LiNGAM results show that retailer has stronger pricing power than CSD manufacturers in most cases.

Obviously, the results of LiNGAM and PC algorithm are quite different. Although all of the estimated error terms reject the normality test, the central question is that whether the error terms are far enough away from the normal distribution to induce a correct estimation in LiNGAM. For this question, Hyvärinen et al. suggest that when measuring the accuracy of the estimation, bootstrapping method should be applied rather than testing the normality (Hyvärinen, Zhang, Shimizu, and Hoyer 2010). This suggestion is thus left for an additional work on causal inference under non-Gaussian data.

**CHAPTER IV**  
**STRUCTURAL DEMAND MODEL FOR THE U.S. CARBONATED SOFT**  
**DRINK MARKET**

**Introduction**

The neoclassical demand system estimation, derived from constrained utility maximization, which assumes that quantity is a function of prices and income and “imposes restriction based on economic theory such adding-up, symmetry, and homogeneity (Perloff, Karp, and Golan, 2007).” These models, including the Almost Ideal Demand System (AIDS) and the Rotterdam demand model, use “flexible functional forms”; that is they leave the own-price and cross-price elasticities evaluated by the data itself without imposing additional assumptions on substitution patterns like Independence of Irrelevant Alternatives (Hausman, 1994). Nonetheless, if the prices of all pertinent goods of differentiated products are considered in neoclassical demand system, estimation of all parameters will be a computational burden, which is well-known as the curse of dimensionality (Pofahl, 2006). Arguably the direct solutions addressing the dimensionality problem include Distance Metric demand estimation method (DM) and discrete choice model (DC).

The discrete choice (DC) demand model reduces the number of coefficients by projecting the number of products on to the number of products’ attribute. It also presumes that each agent only purchases a single unit of the good with the highest utility among all choices (Rojas and Peterson, 2008; Giacomo, 2004). Suppose random-



coefficients logit models represent the theoretical framework for the DC approach. The utility function can be decomposed into a deterministic part and a stochastic part. The researcher is supposed to know the deterministic part, which consists of the observable consumers' preference and the products' attributes, but the distributional shape of the stochastic term should be made (Giacomo, 2004; Nevo, 2000).

The advantages of Random-Coefficients Logit Models are:

(i) The relatively small number of demand parameters has to be calculated by working on the characteristics space. The difference among the consumers' preference is modeled completely and substitution patterns are free to evaluate.

(ii) It is easier to predict consumers' reaction when new brands are introduced to the market. In a neoclassical demand system, if a new product enters the market, all cross-elasticities need to be re-evaluated. However, in a discrete choice model, the new product's market share and elasticities can be investigated without re-estimating the whole demand systems (Giacomo, 2004).

However, one of DC's obvious drawbacks is its restriction on single-unit purchase behavior so consumers' multiple-unit purchases cannot be studied (Giacomo, 2004). Apparently, this assumption does not fit consumer's behavior when buying Carbonated Soft Drink (CSD). Dube (2004) indicated, in his studies, approximately 31% of the shopping trips are multiple-product purchase of CSD and 61.5% of the trips are multiple-unit purchase. It is clear that presumption of single unit purchase is inappropriate in the CSD industry.

On the other hand, the distance metric (DM) estimation method, developed by Pinkse, Slade and Brett (2002), solves the dimensionality problem in neoclassical demand models and the single purchase restriction on DC models. The DM approach specifies the cross-price coefficients semi-parametrically as functions of the distance between the products attribute space. In other words, this approach projects prices dimension into attributes dimension to reduce the number of estimated coefficients (Pofahl, 2006). For instance, in the CSD market, the distance metric  $|Carb_i - Carb_j|$  could be employed, giving the absolute distance of the carbohydrate contents between  $i$  and  $j$ .

In this chapter, I incorporate the Distance Metric (DM) estimation approach into Linear Approximate Almost Ideal Demand System Model (LA/AIDS) to assess the demand after the Cadbury and Dr Pepper/Seven-Up merger which was effective on March 2, 1995 in the U.S. CSD industry.

### **The Carbonated Soft Drink Industry in the United States**

Soft drink production is the largest beverage manufacturing in the U.S., with annual 2006 revenue of \$42.3 billion. This industry is dominated by carbonated soft drinks (CSD), which account for around 54.3% of industry revenue.<sup>22</sup> In 2006, Coca-

---

<sup>22</sup> Soft Drink Production in the US:31211. IBISWorld Industry Report. Online Edition. Available from <http://www.ibisworld.com/industry/default.aspx?indid=284>. Accessed October, 2008.

Cola Company had 42.9% of the CSD market; Pepsi-Cola Company held a 31.2% share, and Cadbury Schweppes owned 14.9% of the market share.<sup>23</sup>

In 1986, PepsiCo planned to buy the Seven-Up Company from Philip Morris, while Coca-Cola was attempting to purchase Dr Pepper. Nevertheless, both proposed acquisitions were rejected by Federal Trade Commission. Later, the investment bank Hicks& Haas purchased both Dr Pepper and the U.S. operations of Seven-Up. Cadbury Schweppes joined in Hicks& Haas's buyout of Dr Pepper and held a minor stake of Dr Pepper. Dr Pepper and Seven-Up are merged to form the Dr Pepper/Seven-Up Companies, Inc. (DPSU) on May 19, 1988.

Since Cadbury sought to become a significant producer of noncola soft drinks; after its acquisition of A&W Brands Inc, their next target is to take over the DPSU. On March 2, 1995, Cadbury Schweppes acquired the rest of DPSU and the new company is called Dr Pepper/Cadbury of North America, Inc. The new company was ranked the third CSD manufacturer with a market share of 17 percent in the U.S. market.<sup>24</sup> Afterwards, Coca-Cola, PepsiCo, and Cadbury together now account for about 90% of all CSDs sold in the U.S. (Saltzman, Levy and Hilke, 1999).

---

<sup>23</sup> Soft Drinks and Bottled Water. Encyclopedia of Global Industries. Online Edition. Gale, 2009. Available from <http://galenet.galegroup.com>. Accessed October, 2008.

<sup>24</sup> Dr Pepper/Seven Up, Inc. Business & Company Resource Center. International Directory of Company Histories, Vol. 32. St. James Press, 2000. Available from <http://galenet.galegroup.com>. Accessed November, 2008.

## Quantitative Methods

### The Demand Model

Pinkse and Slade (2004) derived the aggregate-demand function of product sales according to “a normalized-quadratic indirect-utility function” as

$$q_i = a_i + \sum_j b_{ij} p_j - e_i y + u_i \quad (i=1, \dots, n). \quad (111)$$

where  $B = [b_{ij}]$  is an  $n \times n$  symmetric, negative-semidefinite matrix, and the relevant products' prices  $p = (p_1, p_2, \dots, p_n)^T$  and aggregate income  $y$  are normalized depending on dividing by the outside good's price. They assume that both intercept  $a_i$  and the diagonal elements of  $B$  are functions of the product  $i$ 's attributes,  $x_i$ . That is  $a_i = a(x_i)$  and  $b_{ii} = b(x_i)$ . The off-diagonal elements of  $B$  are supposed to be functions of the distance between some set of products' characteristics,  $b_{ij} = g(d_{ij})$ ,  $i \neq j$ . The function  $g(\cdot)$  is evaluated semi-parametrically rather than giving a fixed form on that, indicating how the distance measures,  $d_{ij}$ , affect the strength of competition between products  $i$  and  $j$ .  $d_{ij}$  measures the closeness of the two products,  $i$  and  $j$ , in attributes space (Pinkse, Slade and Brett, 2002). For example, if the products were brands of bottled juice, the attributes of products might be sodium content, flavor, or dummy variables that indicate whether commodities belong to the same manufacturer. The error term  $u_i$  is mean independent of the observed products characteristics,  $E[u_i | x] = 0$ . If this assumption is violated, the estimator of the parameter of equation (111) is inconsistent.

Rojas (2005) and Pofahl (2006) incorporated the DM approach into the LA/AIDS. The substantial advantages of this model are that it can accommodate the non-linear aggregation over consumers and set no restrictions on the length of the panel data. Pinkse and Slade's individual indirect-utility function is a kind of Gorman polar form. Although it can be easily aggregated or differentiated to obtain brand-level demands, the problematic assumptions are that the change in an individual's demand for certain commodity with respect to a difference in personal income does not depend on earnings; this condition is the same for every consumer regardless of the individual's character. As a result, if a consumer does not buy a product, then the income effect for that product is assumed to be zero. Thus, it amounts to suppose that income effect for all products is zero since it would be simple to find one person who does not purchase a certain commodity, especially with long length time periods in a dataset (Rojas, 2005).

Formally, let  $i \in (1, \dots, N)$  be the index of products,  $t \in (1, \dots, T)$  the set of markets which are defined as cluster-week pairs,<sup>25</sup>  $p_t = (p_{1t}, \dots, p_{Nt})$  the vector of retail prices in market  $t$ ,  $q_t = (q_{1t}, \dots, q_{Nt})$  the vector of quantities demanded, and  $X_t = \sum_i p_{it} q_{it}$  total expenditures in market  $t$ . Using these notations, the LA/AIDS suggested by Deaton and Muellbauer (1980) is given as follows:

$$w_{it} = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln(p_{jt}) + \beta_i \ln\left(\frac{X_t}{P_t^*}\right) + \varepsilon_{it} \quad (112)$$

---

<sup>25</sup> The cross-price elasticities are zero through markets.

where  $w_{it} = \frac{p_{it}q_{it}}{X_t}$  is the expenditure share for product  $i$  in market  $t$ , and the Stone price

index is defined as follows:

$$\ln(P_t^*) \equiv \sum_{i=1}^N w_{it} \ln(p_{it}), \quad (113)$$

It was typical to use Stone price index to linearize the AIDS model. However, Moschini (1995) indicated that Stone index, varies with the variation in units of measurement of prices and quantities. For instance, suppose we change the unit of the first good from bales to tons, then the corresponding price will be scaled by 4 (1 ton = 4 bales). Since such alternation does not impact the expenditure shares, the Stone index would apply unchanged weights to the scaled prices. This problem makes  $\gamma_{ij}$  or  $\beta_i$  generally biased. Moschini suggested one feasible choice:  $\ln(P_t^L)$  which is a loglinear analogue of the Laspeyres index and defined as:

$$\ln(P_t^*) \approx \ln(P_t^L) = \sum_i w_i^0 \ln(p_{it}), \quad (114)$$

where  $w_i^0$  is product  $i$ 's 'base' share, defined as  $w_i^0 \equiv T^{-1} \sum_t w_{it}$ , the average expenditure share of product  $i$  over  $t$ .

After replacing (113) by (114), the sales share form of LA/AIDS can be written as

$$w_{it} = \alpha_i + \sum_{j=1}^N \gamma_{ij} \ln(p_{jt}) + \beta_i \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{it} \quad (115)$$

Normally, the  $(N-1)$  equations of (115) can be estimated by Seemingly Unrelated Regression method. Nevertheless, if we apply LA/AIDS model to assess the demand of

numerous CSD products here, the procedure has a significantly challenge in its evaluation due to the curse of dimensionality.

### **Distance Metric Approach**

Following Rojas (2005) and Pofahl (2006), the distance metrics are added to the LA/AIDS model to alleviate the difficulty of estimation. Briefly, the objective function, (116), includes a vector,  $d$ , which is the distance measure of products' attributes, and the cross-price coefficients,  $\gamma_{ij}$ ,  $i \neq j$ , can be prescribed as a function  $g(\cdot)$  of the distance measures,  $d_{ij}$ .

$$w_{it} = \alpha_i + \gamma_{ii} \ln(p_{it}) + \sum_{j \neq i} g(d_{ij}^k; \lambda) \ln(p_{jt}) + \beta_i \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{it} \quad (116)$$

where  $k$  equals the number of distance measures, and  $\lambda$  is the corresponding coefficients to each distance metric (Pofahl, 2006). The element of  $d$  is determined by researcher. Inwardly, the function  $g$  shows how difference of attributes affects the strength of product's competition (Pinkse, Slade and Brett, 2002).

The own-price parameter  $\gamma_{ii}$  is comprised of a constant and product  $i$ 's attributes. Suppose carbohydrate content is a relevant attribute that has impact on the demand of CSD,  $\gamma_{ii}$  can be written as  $\gamma_{ii} = \gamma_0 + \gamma_1 Carb_i$ ; hence,  $\gamma_{ii} \ln(p_{it}) = \gamma_0 \ln(p_{it}) + \gamma_1 \ln(p_{it}) Carb_i$ , including a price interacting term with the product characteristics.

To have a more clear insight of DM method, let's make a simple example below. Suppose there are four commodities sold in the market, the traditional AIDS demand system is written as:

$$\begin{aligned}
w_{1t} &= \alpha_1 + \gamma_{11} \ln(p_{1t}) + \gamma_{12} \ln(p_{2t}) + \gamma_{13} \ln(p_{3t}) + \gamma_{14} \ln(p_{4t}) + \beta_1 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{1t} \\
w_{2t} &= \alpha_2 + \gamma_{21} \ln(p_{1t}) + \gamma_{22} \ln(p_{2t}) + \gamma_{23} \ln(p_{3t}) + \gamma_{24} \ln(p_{4t}) + \beta_2 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{2t} \\
w_{3t} &= \alpha_3 + \gamma_{31} \ln(p_{1t}) + \gamma_{32} \ln(p_{2t}) + \gamma_{33} \ln(p_{3t}) + \gamma_{34} \ln(p_{4t}) + \beta_3 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{3t} \\
w_{4t} &= \alpha_4 + \gamma_{41} \ln(p_{1t}) + \gamma_{42} \ln(p_{2t}) + \gamma_{43} \ln(p_{3t}) + \gamma_{44} \ln(p_{4t}) + \beta_4 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{4t},
\end{aligned}$$

If we impose symmetry on the cross-price parameters, six cross-price parameters need to be estimated. Suppose carbohydrates and sodium contents have influence on the choice of CSD purchase and let  $d_{ij}^{carb}$  and  $d_{ij}^{so}$  symbolize the distance measures of carbohydrates as well as sodium content between brand  $i$  and  $j$ . The distance metric function for cross-price coefficients can be presumed as  $\lambda_0 + \lambda_1 d_{ij}^{carb} + \lambda_2 d_{ij}^{so}$ . Given there are no brand attributes terms in the own-price parameter, the whole system after substitution becomes

$$\begin{aligned}
w_{1t} &= \alpha_1 + \gamma_0 \ln(p_{1t}) + [\lambda_0 + \lambda_1 d_{12}^{carb} + \lambda_2 d_{12}^{so}] \ln(p_{2t}) + [\lambda_0 + \lambda_1 d_{13}^{carb} + \lambda_2 d_{13}^{so}] \ln(p_{3t}) + \\
&[\lambda_0 + \lambda_1 d_{14}^{carb} + \lambda_2 d_{14}^{so}] \ln(p_{4t}) + \beta_1 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{1t} \\
w_{2t} &= \alpha_2 + [\lambda_0 + \lambda_1 d_{21}^{carb} + \lambda_2 d_{21}^{so}] \ln(p_{1t}) + \gamma_0 \ln(p_{2t}) + [\lambda_0 + \lambda_1 d_{23}^{carb} + \lambda_2 d_{23}^{so}] \ln(p_{3t}) + \\
&[\lambda_0 + \lambda_1 d_{24}^{carb} + \lambda_2 d_{24}^{so}] \ln(p_{4t}) + \beta_2 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{2t} \\
w_{3t} &= \alpha_3 + [\lambda_0 + \lambda_1 d_{31}^{carb} + \lambda_2 d_{31}^{so}] \ln(p_{1t}) + [\lambda_0 + \lambda_1 d_{32}^{carb} + \lambda_2 d_{32}^{so}] \ln(p_{2t}) + \gamma_0 \ln(p_{3t}) + \\
&[\lambda_0 + \lambda_1 d_{34}^{carb} + \lambda_2 d_{34}^{so}] \ln(p_{4t}) + \beta_3 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{3t} \\
w_{4t} &= \alpha_4 + [\lambda_0 + \lambda_1 d_{41}^{carb} + \lambda_2 d_{41}^{so}] \ln(p_{1t}) + [\lambda_0 + \lambda_1 d_{42}^{carb} + \lambda_2 d_{42}^{so}] \ln(p_{2t}) + \\
&[\lambda_0 + \lambda_1 d_{43}^{carb} + \lambda_2 d_{43}^{so}] \ln(p_{3t}) + \gamma_0 \ln(p_{4t}) + \beta_4 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{4t},
\end{aligned}$$

Moreover, the demand system can be written as



$$\begin{aligned}
w_{1t} &= \alpha_1 + \gamma_0 \ln(p_{1t}) + \lambda_0 [\ln(p_{2t}) + \ln(p_{3t}) + \ln(p_{4t})] + \lambda_1 [d_{12}^{carb} \ln(p_{2t}) + d_{13}^{carb} \ln(p_{3t}) + d_{14}^{carb} \ln(p_{4t})] \\
&+ \lambda_2 [d_{12}^{so} \ln(p_{2t}) + d_{13}^{so} \ln(p_{3t}) + d_{14}^{so} \ln(p_{4t})] + \beta_1 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{1t} \\
w_{2t} &= \alpha_2 + \gamma_0 \ln(p_{2t}) + \lambda_0 [\ln(p_{1t}) + \ln(p_{3t}) + \ln(p_{4t})] + \lambda_1 [d_{21}^{carb} \ln(p_{1t}) + d_{23}^{carb} \ln(p_{3t}) + d_{24}^{carb} \ln(p_{4t})] \\
&+ \lambda_2 [d_{21}^{so} \ln(p_{1t}) + d_{23}^{so} \ln(p_{3t}) + d_{24}^{so} \ln(p_{4t})] + \beta_2 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{2t} \\
w_{3t} &= \alpha_3 + \gamma_0 \ln(p_{3t}) + \lambda_0 [\ln(p_{1t}) + \ln(p_{2t}) + \ln(p_{4t})] + \lambda_1 [d_{31}^{carb} \ln(p_{1t}) + d_{32}^{carb} \ln(p_{2t}) + d_{34}^{carb} \ln(p_{4t})] \\
&+ \lambda_2 [d_{31}^{so} \ln(p_{1t}) + d_{32}^{so} \ln(p_{2t}) + d_{34}^{so} \ln(p_{4t})] + \beta_3 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{3t} \\
w_{4t} &= \alpha_4 + \gamma_0 \ln(p_{4t}) + \lambda_0 [\ln(p_{1t}) + \ln(p_{2t}) + \ln(p_{3t})] + \lambda_1 [d_{41}^{carb} \ln(p_{1t}) + d_{42}^{carb} \ln(p_{2t}) + d_{43}^{carb} \ln(p_{3t})] \\
&+ \lambda_2 [d_{41}^{so} \ln(p_{1t}) + d_{42}^{so} \ln(p_{2t}) + d_{43}^{so} \ln(p_{3t})] + \beta_4 \ln\left(\frac{X_t}{P_t^L}\right) + \varepsilon_{4t}
\end{aligned}
\tag{117}$$

For cross-price coefficients, only three parameters,  $\lambda_0$ ,  $\lambda_1$  and  $\lambda_2$  are necessary to be estimated right now. Obviously, if there are numerous products are involved in the demand estimation, the method's effect on reducing the dimensionality will be clearer. This is why it is said DM method can handle the challenge for “curse of dimensionality.”<sup>26</sup>

Because the distance metrics are symmetric, symmetry can be required by making  $\lambda$  equal across all equations. Once we obtain the estimated coefficients of  $\lambda$ , it is simple to calculate the cross-price coefficients and elasticities. The functional form of  $g(\cdot)$  can be estimated by parametric or semiparametric methods. If the parametric assumption is correct, then choosing the semi-parametric methods will be inefficient. However, we estimate  $g(\cdot)$  semi-parametrically since it can derive as much flexibility in

---

<sup>26</sup> This example was suggested by Dr. Pofahl.

the pattern of substitution as possible without depend on any arbitrary parametric form according to the analyst's uncertain knowledge or beliefs

Additionally, the expression of Marshallian price elasticities, uncompensated price elasticities, can be written as (Green and Alston, 1990; Rojas, 2005)

$$\eta_{ij} = \begin{cases} -1 + \frac{1}{w_{it}} \left[ \gamma_{ii} - \beta_i \frac{d \ln P_t^L}{d \ln p_{it}} \right] = -1 + \frac{1}{w_{it}} [\gamma_{ii} - \beta_i w_i^0], & \text{for } i = j \\ \frac{1}{w_{it}} \left[ g(d_{ij}^k; \lambda) - \beta_i \frac{d \ln P_t^L}{d \ln p_{jt}} \right] = \frac{1}{w_{it}} [g(d_{ij}^k; \lambda) - \beta_i w_j^0], & \text{for } i \neq j \end{cases} \quad (118)$$

### Household Demographics

In addition to demand parameter estimation, capturing variation in substitution patterns among across broad classifications of consumer groups should not be ignored in our estimation. Given the same data resource, Hoch, Kim, Montgomery, and Rossi (1995) indicate category-level consumer price elasticity across stores is mainly influenced by consumer demographic difference. In other words, zone-pricing through store chain is substantially activated by price discrimination on consumer's heterogeneity rather than competition between stores. Thus, demographic difference between clusters will be taken into consideration when estimating the demand system of CSD.

## Data and Preliminary Data Statistics

### Scanner Data

The primary data resource is the administrative Dominick Database from the Kilts Center for Marketing at the University of Chicago's Booth School of Business. The

dataset contains weekly retail price, sold quantities, and profits for more than 3500 UPCs for over 100 stores operated by Dominick's Finer Foods (DFF) across approximately 9 years (09/14/89-05/14/97). I select the products related to the Cadbury/DPSU merger from the Soft Drinks group.

### **Data Preparation**

Observations were dropped if one of these variables including sold quantity, price, or profit, was missing. The nearest two years of pre Cadbury/DPSU merger data starting from 03/04/1993 to 03/01/1995, equivalent to recorded week #182 - #285 in DFF database, is employed to simulate the price effects of this CSD merger, while the post-merger data is used to be a comparison with the simulated price changes. Nevertheless, we drop the observations in week #211 because of a lack of available price information for three brands on the selected list. Given selected brands, if a certain brand does not have sales information in some markets during the period of interest, the whole selected brands observations are deleted in that cluster-week pair to avoid bias in estimation. Finally, 5,616 observations are taken into considerations in the analysis.

The information on nutritional facts is based on previous research, as well as collected from CSD package at local supermarkets or manufacturer websites. If the collected information from grocery stores is different from Dube (2005)<sup>27</sup> and McMillan's paper (2007), we will pick up the data from their search rather than information on package since some brands may have been re-formulated. For example, 7 Up has replaced sodium citrate with potassium citrate to reduce the beverage's sodium

---

<sup>27</sup> The first version of Dube's paper (2004) was received on September 25, 2000 by Marketing Science. The products' characteristics table in that paper is the earliest attribution information I can find.

content in 2006.<sup>28</sup> For some commodities which are unavailable through their inquiries, unless we are very sure some brands' formula was altered after 1995, we presume that the nutrition facts on current brands are consistent with the characteristic content during the related period of our research. Another difficulty confronted in this research is that the element of CSD is treated as a business secret; consumer service employees of these companies refused to offer information about ingredients. Initially, different package sizes of specific brand are viewed as different products because of the differences in storability. Also, Dube's research outcome (2005) shows that package size is a relevant attribute affecting consumer's purchase behavior. However, the results of our preliminary Ordinary least squares (OLS) regressions before aggregation indicate package size does not have significant power affecting consumer's purchase behavior as well as shoppers have strong brand loyalty<sup>29</sup>. Besides, most papers applying DM method as Rojas' work in brewing industry and Pofahl's job in bottled juice category mainly focus on brand-level demand. Thus, we do the aggregate of CSD products into sixteen brands and that should be helpful to simply the demand estimation effectively.

---

<sup>28</sup> See [http://www.solarnavigator.net/solar\\_cola/7up.htm](http://www.solarnavigator.net/solar_cola/7up.htm).

<sup>29</sup> Preliminary OLS outcomes are shown in Table F1 and F2 in Appendix F. The t-statistic of package size (Mvol) is -0.90 in Table F1 where the package size is used to set continuous metrics while the t-statistic of package size (Msize) is 2.09 in Table F2 where package size is treated as a dummy indicator. Both of them are insignificant at 5% level.

**TABLE 4 General Traits of Typical Household in Demographic Cluster in DFF Database**

Traits	Cluster			
	A	B	C	D
Description	Established Suburban Families	City Dwellers	Ethnic Neighborhoods	Prospering Suburban Families
Household Size	Medium	Small	Medium	Large
Married	Married (50% w/ children)	Few married	Married	Nuclear Families
Children	Older (6-17)	Few	Few	Many
Singles	Few	Lots	Few	Few
Education	High (36% college+)	Medium (30% college +)	Low education	High (35% college+)
Seniors	Some	Some	Many	Few
Middle Age	Lots	Few	Lots	Few
Dual Income	Lots	Few	Few	Many
Income	Higher (45% \$50000+)	Lower (42% \$20000-)	Lower-middle (80% \$50000-)	Higher (44% \$50000+)
Price Zone	Moderate competition	Low competition	Moderate	Very competitive
Ethnicity		Substantial Blacks, Hispanics		

Notes: We thank William Minseuk Cha., the research assistant of James M. Kilts Center for Marketing, offering this table to us.

\* Nuclear Family primarily refers a family group is comprised of most naturally, a father, a mother and their kids.

DFF clusters its stores into four groups: A, B, C and D. The stores within the same cluster as formed the same cohort and these groups are viewed as separate regional markets. The demographics description and summary of statistics for each cluster's demographic variables are shown in Table 4 and Table 5. This ZIP code level demographic data are obtained from US Government (1990) census data for the Chicago metropolitan area and processed by Market Metrics to generate demographic profiles for

each of the DFF stores.<sup>30</sup> Although it has been documented that Dominick's price zones are up to 16 for the whole Chicago area, for simplicity, the stores are classified into four clusters in this study.

To further simplify the analysis, CSDs aggregated by their brands of all the bottle size with at least a 1% sales volume share (in fluid ounces)<sup>31</sup> of whole market consumption volumes are considered. Finally, it contains 13 brands, including the top-10 soft drinks brands in 1994-1995 and some Cadbury's famous brands; representing approximately 68.3% share of total CSD sales by dollar value during the relevant time period.<sup>32</sup> Chosen brands and their market shares as well as their characteristics are shown in Table F4 and Table F5. Coke is the most expensive while Diet A&W Root Beer is the least expensive. I do not consider regional brands here because of the comparably small nationwide sales percentage, about 3%.

### **Distance Metrics**

The brand attributes that are presumed to affect consumer's perception are comprised of: calories, milligrams of sodium, and grams of total carbohydrates content based on per 12 fluid ounce (355 ml) serving, as well as a set of binary variables for the presence of caffeine, citric acid and whether it is a cola drink. Dummy variables are constructed to identify different manufacturers. These chosen characteristics are established based on earlier work of Dube (2004 and 2005) and McMillan (2007).

---

<sup>30</sup> See <http://research.chicagogsb.edu/marketing/databases/dominicks/demo.aspx> for more detail of DFF Store-Specific Demographics database.

<sup>31</sup> In highly competitive differentiated industry, a new product with 0.5% market share can be considered quite successful (Cotterill, 1999).

<sup>32</sup> The ranking of top-10 best selling brands has a slight change in 1994-1995 but the list of brand for 1994 and 1995 are the same. Database: Business Source Complete.

Noticeably, it is clear that there is high correlation between calories and carbohydrates and therefore we only choose carbohydrates, sparing calories, in setting distance matrices.

Coverage, the percentage of stores that sell specific brand, is utilized as a choice of continuous variable in both Pinkse and Slade's (2004) as well as Rojas's (2005) research. We do not consider it here since almost all of the selected CSDs are sold at every chain store over the interested time period that makes coverage useless here.

Discrete and continuous matrices are set as an inverse of distance to make the interpretation of result easier.

#### *Continuous Distance Measures with Continuous Brand Attributes*

I create single-dimension distance metrics of carbohydrate content of CSD in continuous attribute space as:

$$d_{ij}^{carb} = \frac{1}{1 + 2|Carb_i - Carb_j|} \quad (119)$$

where  $|Carb_i - Carb_j|$  is the absolute value for the difference of brand's rescaled-carbohydrates content and  $d_{ij}^k \in (0,1]$ . If brands  $i$  and  $j$  have the same carbohydrates attributes, this metric reaches the maximized value of 1. As the distance in carbohydrates space between brands  $i$  and  $j$  grows, the metric's value approaches to zero. Obviously, the assumption behind this formula is that the strength of the competition is influenced by how near the brand's attributes are. That is to say that we use this measure to examine if Diet Pepsi is a stronger substitution for Diet Coke than Coke.

**TABLE 5 Statistics of Demographics for Store Cluster in DFF Database**

Variable	Description	Cluster			
		A	B	C	D
Educ	Decimal of College Graduates	0.2716 (0.1066)	0.1970 (0.1141)	0.1303 (0.0630)	0.2686 (0.0962)
Ethnic	Percentage of Blacks & Hispanic population	0.0573 (0.029)	0.4851 (0.2660)	0.1547 (0.1229)	0.0942 (0.0813)
Hval150	Decimal of Households with Value over \$150,000	0.4832 (0.2132)	0.2957 (0.2494)	0.1495 (0.1710)	0.3878 (0.2124)
Hsizeavg	Average Household Size	2.6330 (0.0960)	2.3993 (0.4230)	2.6297 (0.1515)	2.8337 (0.2059)
Income	Log of Median Income	10.7886 (0.1833)	10.1291 (0.15)	10.4911 (0.1078)	10.7767 (0.1731)
Poverty	Decimal of Population with income under \$15,000	0.0329 (0.0117)	0.1425 (0.0425)	0.0694 (0.0174)	0.0293 (0.0120)
Single	Decimal of Singles	0.2436 (0.0205)	0.4127 (0.0714)	0.2716 (0.0217)	0.2526 (0.0258)

Source: DFF database, James M. Kilts Center, University of Chicago Booth School of Business.



The measures of the other continuous characteristics, such as sodium as well as carbohydrates coverage, are constructed by applying the same formula. The above one-dimensional metrics are not singular option for making distance measure; that is we can define an n-dimensional Euclidian distance measure to accommodate multiple attributes between different brands. For instance, a two-dimensional distance metrics can be written as:

$$d_{ij}^{SCB} = \frac{1}{1 + 2\sqrt{(So_i - So_j)^2 + (Carb_i - Carb_j)^2}} \quad (120)$$

However, if we want to know which characteristic plays the most influential role in determining patterns of substitution, a single-dimensional metrics cannot be neglected (Pofahl, 2006).

Actually, we can also apply other function forms for one-dimensional distance matrices

$$d_{ij}^{carb} = \frac{1}{1 + 2\sqrt{(Carb_i - Carb_j)^2}} \quad \text{or} \quad d_{ij}^{carb} = \frac{1}{1 + |\log(Carb_i) - \log(Carb_j)|}$$

because functional form is not an important concern for application of data-driven nonparametric estimations. Even if we select different forms or apply alternative nonparametric techniques, the final estimated outcome will be quite closed or equal to each other. For instance, suppose the real distribution pattern of data is  $\sin(x)$ . Bill uses

$y = x$  while Kent chooses  $z = \frac{1}{2}x$  to process evaluation. Obviously, given the same dataset, Bill's estimated solution should be  $g(y) = \sin(y)$  and Kent's outcome will be

$f(z) = \sin\left(\frac{1}{2}(2x)\right) = \sin\left(\frac{1}{2}z\right)$  since both results are determined by the shape of data

itself. Although the interpretation of estimated parameters is unlike, the evaluated functional forms are equivalent inwardly. However, we must impose some constraints on the estimated coefficients or the outcome can be any arbitrary value.<sup>33</sup>

#### *Discrete Distance Measures with Continuous Brand Attributes*

Following Pinkse and Slade (2004), Rojas (2005) as well as Pofahl (2006), continuous commodities attributes can be used to construct two-dimensional market areas and these measures are derived from the Euclidean distance. Two kinds of metrics are considered here: the nearest-neighbor measures and the common-boundary measures.

For nearest-neighbor metrics, the distance measure in sodium/carbohydrates space can be defined exogenously that  $d_{ij}^{NSC}$  equals to one when brands  $i$  as well as  $j$  are nearest neighbors to each other in sodium/carbohydrates space,  $\frac{1}{2}$  if brands  $i(j)$  is  $j$ 's( $i$ 's) nearest neighbor but not vice versa, and 0 otherwise. Brand  $i$ 's nearest neighbor is meant to be the brand having the shortest Euclidean distance from brand  $i$  in relevant attribute space. To derive more reasonable and reliable Euclidean distance between brands, continuous attributes are rescaled through dividing by its maximum value since each of these characteristics' measurement unit differs so it is better to limit the value of continuous characteristics between 0 and 1.

---

<sup>33</sup> This example was suggested by Dr. Qi Li.

Moreover, for common-boundary metrics,  $d_{ij}^{CBS}$  is set to be one when brands  $i$  and  $j$  share a common boundary in brand's sodium/carbohydrates space but are not nearest neighbors, and zero otherwise. In detail, given the coordinates of  $i$  and  $j$  as  $(i_{so}, i_{carb})$  and  $(j_{so}, j_{carb})$  in sodium/carbohydrates space, then a common boundary of  $i$  and  $j$  is defined as a set of variables  $(Sodium, Carb)$  satisfying the next equation:

$$\sqrt{(So - i_{so})^2 + (Carb - i_{carb})^2} = \sqrt{(So - j_{so})^2 + (Carb - j_{carb})^2} \quad (121)$$

After solving (4.3), a linear relation between  $So$  and  $Carb$  is that:

$$So = Carb \frac{i_{carb} - j_{carb}}{j_{so} - i_{so}} + \frac{j_{so}^2 + j_{carb}^2 - i_{so}^2 - i_{carb}^2}{2(j_{so} - i_{so})} \quad (122)$$

Once above equation for all  $i$  and  $j$  are solved, the intersection points of the lines derived from linear equation will be determined and necessarily establish which portion of the lines are actual common boundaries (Rojas, 2005).

Additionally, another set of nearest-neighbor is developed by considering brand attributes and per fluid ounce price together. It allows a situation that consumer's purchase decision depends on both brands' attributes as well as the relative prices between competitors simultaneously (Rojas, 2005). Following Rojas, nearest-neighbor metrics is set upon the summation of square for the attributes' Euclidean distance and differential in average per fl.oz. price. That is,

$$(So - i_{so})^2 + (Carb - i_{carb})^2 + p_i = (So - j_{so})^2 + (Carb - j_{carb})^2 + p_j \quad (123)$$

### *Discrete Distance Measures with Discrete Brand Attributes*

Here, some categories distance measures are included. Although there is no absolute criterion on classification of CSDs, the selected commodities are classified based on their cola or caffeine content. In other words,  $d_{ij}^{cola}$  is equal to one if brands  $i$  as well as  $j$  are cola-type soft drink and zero otherwise. Also,  $d_{ij}^{caff}$  is equal to one if both brands  $i$  as well as  $j$  are caffeine (or caffeine-free) soft drink and zero otherwise

Besides, we have dummy variable indicating if the drinks have ingredient of citric acid and thus  $d_{ij}^{citric}$  is set to be one if brands  $i$  as well as  $j$  both contain citric acid and zero otherwise. These products are manufactured by Coca-Cola, PepsiCo, and Cadbury respectively so a discrete distance metric for manufacturer identity is created to examine if shoppers tend to substitute between brands with the same manufacturer when price change occurs. Hence,  $d_{ij}^{manu}$ 's value is one when brands  $i$ , as well as  $j$ , belong to the same manufacturer and zero otherwise.

All weighting matrices regarding product classification can be normalized; the sum of each row is equivalent to one and thus the weighted prices of rival commodities for the same type will be equal to their mean (Rojas, 2005).

### **Estimated Results**

Even though OLS (or IV) estimated coefficients are probably inconsistent, they are still meaningful. Here, the percentage of stores on sales for specific brand in the same cluster is taken to be a part of intercept when running preliminary OLS regression.

### Preliminary OLS Regression Result of Disaggregation

Prior to doing an aggregate of products by their brands, there are 15,850 observations, consisting of 50 products. Table F1 and F2 shows the estimated coefficients and t-statistics results of each distance measure when package size is treated as continuous metric and discrete metric, respectively. Regardless of the form of container size metrics, the OLS preliminary results indicate the own-price coefficients are negative and statistically significant even at 1% level.<sup>34</sup> The result also indicates that when the rival product's price increases, it stimulates the consumption of own goods.<sup>35</sup> We do not check the level of substitution for a particular product because we cannot trace specific consumer's shopping history in the database and thus it is impossible to build relevant metric for this item. On the other hand, we set a brand identity,  $d_{ij}^{brand}$  to check if there is a stronger substitution between the carbonated soft drinks with the same brand. As expected, the positive coefficient on brand identity shows that a good selling of Coke 6-package will reduce Coke 12-package sales. In addition, both cases indicate the products' promotion activity can boost its sales.<sup>36</sup> However, the estimated result of  $d_{ij}^{manu}$  implies consumers do not have tendency to substitute between the products of the same manufacturer. Finally, we pay attention on the interpretation of group classifications,  $d_{ij}^{cola}$  as well as  $d_{ij}^{caff}$ . The coefficient on coke segment takes positive value while the coefficient on caffeine segment takes negative value. Since brands that

---

<sup>34</sup> The t-statistic of the own-price coefficient is -43.88 when the package size is used to set continuous metrics and -42.04 when package size is treated as a dummy indicator.

<sup>35</sup> The t-statistic of the rival-price coefficient is 3.39 when package size is treated as a dummy indicator

<sup>36</sup> The t-statistic of sales coefficient is 4.02 when the package size is used to set continuous metrics and 3.82 when package size is treated as a dummy indicator.

belong to the same product classification should be substitutes, the negative coefficient for caffeine segment seems to be a wrong sign. Nevertheless, although cola segment has positive coefficient, the estimated value is insignificant. The outcome shows both group classifications are inappropriate. The comparison between our estimation result and Dube's result (2004 and 2005) is shown in Table F3.

### **Preliminary OLS Regression Result after Aggregation**

Results of our preliminary OLS regressions before aggregation indicates package size is not an obviously relevant attribute affecting the purchasing decision.<sup>37</sup>

Consequently, the relevant data of CSD products is aggregated by their brands.

Estimation results are reported in Table 6. Most distance metrics have similar effects as the result of disaggregated products. For example, sales activity can stimulate consumer to purchase and also the competition between products in the same category is more aggressive. Besides, the nearest neighbor measure with price has stronger effect than its counterparts.

### **Empirical Problem and Conclusion**

When the symmetry condition is imposed in the estimation, there is a potential problem, not all own-price coefficients can be less than zero and not all cross-price parameters can be greater than zero.

---

<sup>37</sup> The t-statistic of can size indicator (Msize) is 1.83 which is insignificant under 95% confidence level.

**TABLE 6 OLS Regression Results of Estimated Coefficient on Distance Metrics after Aggregation**

Distance Metrics	Cross-Price	
	Coeff	t-stat.
Continuous Distance Measures with Continuous Variables		
One-Dimensional		
Carbohydrate Content (Mcarb)	11.85*	4.28
Sodium Content (Mso)	-27.52*	-7.00
Two-Dimensional		
Sodium/Carbohydrate Content (MSC)	11.23	1.53
Discrete Distance Measures with Continuous Variables		
Nearest Neighbor		
Sodium/Carbohydrate Content (MNNSC)	-7.74	-1.15
Sodium/Carbohydrate/Price Content (MNNSCP)	24.58*	4.18
Common Boundaries		
Sodium/Carbohydrates Content (MCBSC)	-28.35*	-12.49
Discrete Distance Measures with Discrete Variables		
Product Classifications		
Product grouping (Mgroup)	-46.38*	-5.48
Citric Acid Containing (Mcitric)	6.54	0.43

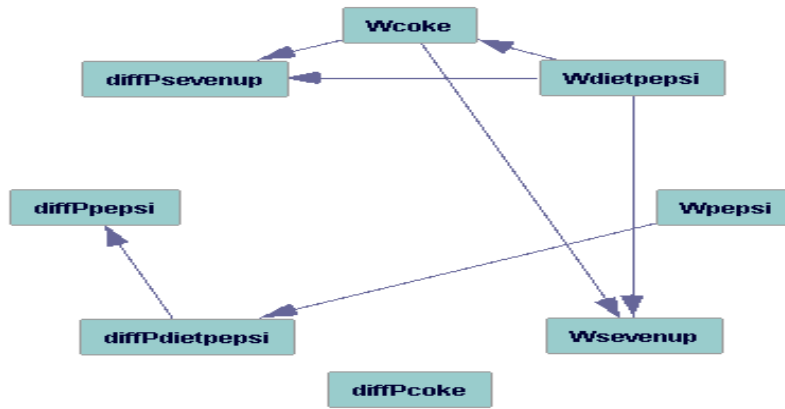
1. All regressions include cluster, product, and year dummy indicators.
2. Coefficients have been multiplied by 1,000 for readability
3. \* Significant at 1%

The possible problem is price endogeneity. Dhar et al. (2003) shows AIDS model with retail level scanner data for differentiated products has price endogeneity, likely coming from retailer's pricing strategy or consumer heterogeneity, and that will cause inconsistent demand estimates as well as have large impact on price and expenditure elasticities. Apparently, the problem of price endogeneity should be dealt with and finding a suitable instrument is always a solution to this problem.

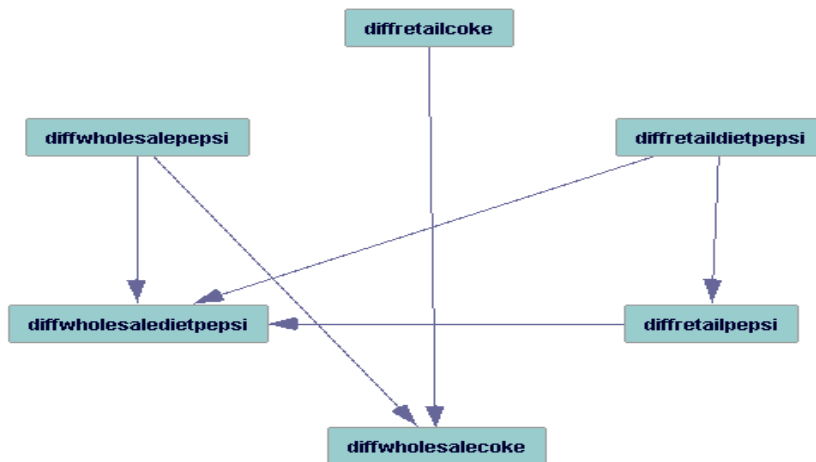
The first-four brands of most selling volume in our dataset are Pepsi, Coke, Diet Pepsi, and 7 Up, so the series of expenditure share and logarithm of these brands' retail prices are examined by LiNGAM algorithm to see their causal patterns. Noticeably, the retail prices series are all non-stationary so we take the first difference of the retail prices series to have stationary series. After taking the first difference of prices series, most variables studied here reject the null hypotheses of symmetry and normality tests and this is why LiNGAM is used in this case. In Figure 4, it shows that there is a causal connection between the retail price of Pepsi and Diet Pepsi; that implies the possibility of price endogeneity. The retail price of Pepsi and Diet Pepsi are associated because of their mutual relation with an unobserved common shock. Pofahl (2006) took the wholesale price of each good to be the instrument of the retail price because he thinks the wholesale price has impact on the setting of retail price of own products but should uncorrelated with the other wholesale prices. Therefore, the causal pattern between the retail price and whole price is searched again to check if the wholesale price can be a proper instrument. All of the retail and wholesale prices are taken first difference to have stationary series.



**FIGURE 4 Causal Connection between the Expenditure Share and First Difference of the Retail Prices Searched by LiNGAM with Prune Factor=0.7<sup>38</sup>**



**FIGURE 5 Causal Connection between the Retail Price and the Wholesale Price Searched by LiNGAM with Prune Factor=0.7<sup>39</sup>**



<sup>38</sup> Wcoke represents the wholesale price of Coke and diffPcoke represents the first difference of the retail price of Coke. Other notations follow the same rules.

<sup>39</sup> diffretailcoke denotes the first difference of Coke's retail price and diffwholesalecoke denotes the first difference of Coke's wholesale price. Other notations follow the same rules.

In Figure 5, the results describe that the retail price has impact on the brand's charge from manufacturer. The brand's manufacturer's price is not only affected by its own retail price but also influenced by other brands' wholesale price. For example, the manufacturer's selling price of Diet Pepsi is not only influenced by its own retail price but also by the wholesale and retail price of Pepsi. As Dhar, Chavas, Coterill and Gould (2005) referred, one manufacturer (e.g. PepsiCo) has several brands sold in one market simultaneously so the manufacturer sales these brand with different strategy. It is not surprising that the wholesale prices of Pepsi and Diet Pepsi are interacted and manufacturer's marketing strategy influences the wholesale prices of its brands mutually. However, this also proves that wholesale price may not be a good instrument for the retail price as Pofahl suggested. It points out the difficulty of finding a proper instrument. Although the OLS results satisfy some expectation in marketing studies, the outcome is biased. It is difficult to have consistent estimates in our case because of the lack of suitable instruments. However, it emphasizes the importance of causal search. Even if the structural modeling is developed on the basis of sound economic theory, economic theory is not correct in every condition and the real world data may operate against the theory. Truth should be built on the evidence so how this world really work should not depend on the researcher's personal knowledge; observing the pattern of real-world data is necessary.

## **CHAPTER V**

### **CONCLUSIONS**

#### **Key Findings**

There exists many statistical methods to evaluate the association between variables and test the significance of these associations. However, the significance of association cannot specify the causal connection between the variables. If we seek a possible result from a given policy or intervention, a causal knowledge is necessary. PC algorithm, based on the research works of Pearl (2009) and Spirtes, Glymour and Scheines (2001), is applied to search for the causal pattern of real-world uncontrolled datasets. However, PC algorithm works more reliably on normally-distributed or symmetrically, but non-normally distributed data. Its performance in discovering causal structures on extremely non-Gaussian dataset is not so good. Also, it is common for PC algorithm to generate a set of unidentifiable directed acyclic graphs (DAGs), especially in the case of fewer observed variables. Shimizu et al. (2006) developed Linear Independent Non-Gaussian Model (LiNGAM) to do causal search based on the independently non-Gaussian distributed disturbances by applying higher-order moment structures. More non-Gaussian data works better in LiNGAM.

These two algorithms are applied to retail-level scanner data in order to investigate the pricing power between manufacturer and retailer in carbonated soft drinks market. PC can only say if the variables are related, but cannot identify the causal direction. LiNGAM gives more exact causal patterns. In general, at least 74% of the products

studied have a pricing pattern such that the retail pricing affects the charge from manufacturer when prune factor is set to be one. If we lower the strength of pruning, the result that retailer has stronger pricing power than the manufacturer is still seen. In addition, the retail price affects the consumers' purchased behavior is also observed, as consumer theory anticipates. Surprisingly, there are several graphs uncovered that show that the manufacturer's price also affects the number of package sold. A possible reason for such a result is the scarcity of shelf space. The price offered from the manufacturer to retailer may have an impact on the product's shelf position and thus may affect the sales condition indirectly. Of course, the real pricing strategic game is very complicated and not well understood, so it is not proper to easily conclude a specific pricing pattern must represent a given strategic game, such as Retailer Stackelberg Leadership or Manufacturer Stackelberg Leadership. On the other hand, the estimation and framework of structural equation model for strategic interactions in a distribution channel is commonly complex. Although our model cannot provide more exact evaluation of gross-margin or marginal operating cost, it does offer a more efficient way to check the interaction between manufacturer and retailer on pricing strategies and give a possible direction for the further research.

In Chapter IV, we incorporate the distance metric into linearized almost ideal demand system to investigate the consumer purchase behavior in carbonated soft drink market. The ordinary least squares estimates give some reasonable outcomes as consumer demand analysts predict, unfortunately, when further restrictions are imposed as the theory suggests, we do not find reasonable results, as most researchers require. It

implies that economic theory does not (in this case) correspond to the movement of real-world data; it also shows the importance of causal inference research.

### **Possible Future Research**

In terms of future research, I would like to investigate the question of pruning and its possible implications and consider simulation to check the accuracy of the LiNGAM estimated results. As Dr. Hoyer and Dr. Shimizu suggested to me, the prune factor is implemented mainly for computational efficiency and it might not have very strong theoretical support. Dr. Shimizu also suggests that it might be better to do bootstrapping to see if the connection between two variables is significant, fixing the variable ordering to be the estimated one by LiNGAM and doing ordinary least squares on the bootstrap samples. Therefore, this is the next step I attempt to do to modify the current results. Besides, LiNGAM works better in more non-normal distributed data. The distribution shape of the number of package sold is quite far from normal distribution but the other two variables, retail price and manufacturer's price, are not extremely departed from normality. Therefore, a simulation is necessary to check the accuracy of our current estimates.

## REFERENCES

- BABA, K., SHIBATA, R., AND SIBUYA, M. "Partial Correlation and Conditional Correlation as Measures of Conditional Independence." *Australian and New Zealand Journal of Statistics*, Vol. 46 (2004), pp. 657-664.
- BENTLER, P.P. "Some Contributions to Efficient Statistics in Structural Models: Specification and Estimation of Moment Structures." *Psychometrika*, Vol. 48 (1983), pp. 493-517.
- BOLLEN, K.A. *Structural Equations with Latent Variables*. 1<sup>st</sup> edition, Hoboken, NJ: Wiley-Interscience, 1989.
- CHOI, S. K., GRANDHI, R., AND CANFIELD, R. A. *Reliability-based Structural Design*. 1<sup>st</sup> edition, New York: Springer, 2006.
- COTTERILL, R. "The Economics of Private Label Pricing and Channel Coordination." In G. Galizzi and L. Venturinin, eds., *Vertical Relationships and Coordination in the Food System*. New York: Physical-Verlag, 1999.
- COVER, T.M., AND THOMAS, J.A. *Elements of Information Theory*. 2<sup>nd</sup> edition, Hoboken, NJ: Wiley-Interscience, 2006.
- DEATON, A., AND MUELLBAUER, J. *Economics and Consumer Behavior*. New York: Cambridge University Press, 1980.
- DHAR, T., CHAVAS, J.P., AND GOULD, B.W. "An Empirical Assessment of Endogeneity Issues in Demand Analysis for Differentiated Products." *American Journal of Agricultural Economics*, Vol. 85, No. 3 (2003), pp. 605-617.
- \_\_\_\_\_, CHAVAS, J.P., COTTERILL, R.W. AND GOULD, B.W. "An Econometric Analysis of Brand-Level Strategic Pricing between Coca-Cola Company and PepsiCo." *Journal of Economics and Management Strategy*, Vol. 14 (2005), pp. 905-931.
- DODGE, Y., AND ROUSSON V. "On Asymmetric Properties of the Correlation Coefficient in the Regression Setting." *The American Statistician*, Vol. 55 (2001), pp. 51-54.
- DUBE, J.P. "Multiple Discreteness and Product Differentiation: Demand for Carbonated Soft Drinks" *Marketing Science*, Vol. 23 (2004), pp. 66-81.
- \_\_\_\_\_. "Product Differentiation and Mergers in the Carbonated Soft Drink Industry" *Journal of Economics & Management Strategy*, Vol. 14 (2005), pp. 879-904

- DOMINGUEZ, F.F. "Inductive Causation on Strategic Behavior: The Case of Retail and Manufacturer Pricing," Ph. D. Dissertation, Department of Agricultural Economics, Texas A&M University, 2009.
- GIACOMO, M.D. "Empirical Analysis of Markets with Differentiated Products: The Characteristics Approach." *Giornale degli Economisti e Annali Economia*, Vol. 63 (2004), pp. 243-288.
- GLYMOUR, CLARK. E-mail to author. 28 September 2010.
- GREEN, R., AND ALSTON, J "Elasticities in AIDS models." *American Journal of Agricultural Economics*, Vol. 72 (1990), pp. 442-445
- HAUSMAN, J. "Valuation of New Goods under Perfect and Imperfect Competition." Working Paper no. 4970, National Bureau of Economic Research, 1994.
- HYVÄRINEN, A. "Fast ICA by a Fixed-point Algorithm that Maximize Non-Gaussianity." In S. Roberts and R. Everson, eds., *Independent Component Analysis: Principles and Practice*, 1<sup>st</sup> edition, New York: Cambridge University Press, 2001.
- \_\_\_\_\_, KARHUNEN J., AND OJA, E. *Independent Component Analysis*. 1<sup>st</sup> edition, Hoboken, NJ: Wiley-Interscience, 2001.
- \_\_\_\_\_, ZHANG, K., SHIMIZU, S., AND HOYER, P.O. "Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity." *Journal of Machine Learning Research*, Vol. 11 (2010), pp. 1709-1731.
- HOCH, S.J., KIM, B.D., MONTGOMERY, A.L., AND ROSSI, P.E. "Determinants of Store-Level Price Elasticity", *Journal of Marketing Research*, Vol. 32 (1995), pp. 17-30.
- HOOVER, K.D. "Causality in Economics and Econometrics." In S. N. Durlauf and L. E. Blume, eds., *The New Palgrave Dictionary of Economics*, 2<sup>nd</sup> Edition. Palgrave Macmillan, 2008.
- \_\_\_\_\_. "Identity, Structure, and Causal Representation in Scientific Models." invited lecture at the conference on *Modeling the World: Perspectives from Biology and Economics*, University of Helsinki, (2009a); an earlier version was posted as "Identity, Structure, and Causation in Scientific Models."
- \_\_\_\_\_. "Probability and Structure in Econometric Models." In C. Glymour, W. Wang and D. Westerstahl, eds., *Logic, Methodology and Philosophy of Science*:

*Proceedings of the Thirteenth International Congress*. London: College Publications, 2009b.

INGENE C.A., AND PARRY M.E. *Mathematical Models of Distribution Channel*. 1<sup>st</sup> edition, New York: Springer, 2004.

KADIYALI, V., CHINTAGUNTA P., AND VILCASSIM N., “Manufacturer-Retailer Channel Interactions and Implications for Channel Power: An Empirical Investigation of Pricing in a Local Market.”, *Marketing Science*, Vol. 19 (2000), pp. 127-148.

KANO, Y., AND SHIMIZU, S., “Causal Inference Using Nonnormality.” In T. Higuchi, Y. Iba, AND M. Ishiguro, eds., *Proceedings of the International Symposium on Science of Modeling-the Thirtieth Anniversary of the Information Criterion (AIC), ISM Report on Research and Education, No. 17*. Tokyo: The Institute of Statistical Mathematics, 2003.

LACERDA, G., SPIRITES, P., RAMSEY J., AND HOYER, P.O. “Discovering Cyclic Causal Models by Independent Components Analysis.” *Proceedings of the Twenty-fourth Conference on Uncertainty in Artificial Intelligence*, Helsinki, Finland, 2008.

LEON, S.J. *Linear Algebra with Application*. 6<sup>th</sup> edition, New Jersey: Prentice Hall, 2006.

MCMILLAN, R.S. “Different Favor, Same Price: The Puzzle of Uniform Pricing for Differentiated Products?” *Federal Trade Commission, Bureau of Economics*, (Jan., 2007)

MEEK, C. “Causal Inference and Causal Explanation with Background Knowledge.” In P. Besnard and S. Hanks, eds., *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers, Inc., 1995, pp. 403-441.

MILLER, J.E. *The Chicago Guide to Writing about Multivariate Analysis*. University of Chicago Express, 2005.

MOOIJART, A. “Factor Analysis for Non-Normal Variables.” *Psychometrika*, Vol. 50 (1985), pp.323-342.

MOSCHINI, G. “Units of Measurement and the Stone Index in Demand System Estimation.” *American Journal of Agricultural Economics*, Vol. 77 (1995), pp. 63-68

NEVO, A. “Mergers with Differentiated Products: The Case of the Ready-to-eat Cereal Industry.” *RAND Journal of Economics*, Vol. 31 (2000), pp.395-421.



- PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1<sup>st</sup> edition, San Francisco: Morgan Kaufmann, 1988.
- \_\_\_\_\_. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press, 2<sup>nd</sup> edition, 2009.
- PERLOFF, J.M., KARP L.S., AND GOLAN, A. *Estimating Market Power and Strategies*. New York: Cambridge University Press, 2007.
- PINKSE, J., AND SLADE, M.E. “Mergers, Brand Competition and the Price of a Pint.” *European Economic Review*, Vol. 48 (2004), pp. 617-643.
- \_\_\_\_\_, SLADE, M.E., AND BRETT, C. “Spatial Price Competition: A Semiparametric Approach.” *Econometrica* 70 (2002), pp. 1111-1153.
- POFAHL, G. “Essays on Horizontal Merger Simulation: the Curse of Dimensionality, Retail Price Discrimination, and Supply Channel Stage-Games,” Ph. D. Dissertation, Department of Agricultural Economics, Texas A&M University, 2006.
- RICHARDSON, T., AND SPIRITES, P. “Ancestral Graph Markov Models.” *The Annals of Statistics* Vol. 30 (2002), pp. 962-1030.
- ROJAS, C. “Demand Estimation with Differentiated Products: An Application to Price Competition in the U.S. Brewing Industry,” Ph. D. Dissertation, Department of Economics, Virginia Polytechnic Institute and State University, 2005.
- \_\_\_\_\_. “Price Competition in U.S. Brewing.” *Journal of Industrial Economics*, Vol. 56 (2008), pp.1-31.
- \_\_\_\_\_, AND PETERSON E.B. “Demand for Differentiated Products: Price and Advertising Evidence from the U.S. beer Market.” *International Journal of Industrial Organization*, Vol. 26 (2008), pp. 288-307.
- SALTZMAN, H., LEVY, R., AND HILKE C., JOHN “Transformation and Continuity: the U.S. Carbonated Soft Drink Bottling Industry and Antitrust Policy Since 1980.” *Bureau of Economics Staff Report, Federal Trade Commission*, November 1999.
- SEARLE, S.R., AND WILLETT, L.S. *Matrix Algebra for Applied Econometrics*. 1<sup>st</sup> edition, Hoboken, NJ: Wiley-Interscience, 2001.
- SHIMIZU, S., HOYER P.O., HYVÄRINEN, A. AND KERMINEN, A. “A Linear Non-Gaussian Acyclic Model for Causal Discovery.” *Journal of Machine Learning Research*, Vol. 7 (2006), pp. 2003-2030.

- \_\_\_\_\_, HYVÄRINEN, A., KANO, Y., AND HOYER, P.O. “Discovery of Non-Gaussian Linear Causal Models Using ICA.” *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, Quebec, Canada pp. 526-533, 2005.
- \_\_\_\_\_, HYVÄRINEN, A., HOYER P.O., AND KANO, Y. “Finding a Causal Ordering via Independent Component Analysis.” *Computational Statistics and Data Analysis*, Vol. 50 (2006), pp. 3278-3293.
- \_\_\_\_\_, AND KANO, Y. “Use of Non-Normality in Structural Equation Modeling: Application to Direction of Causation.” *Journal of Statistical Planning and Inference*, Vol. 138 (2008), pp. 3483-3491.
- \_\_\_\_\_. E-mail to author. 16 September 2010.
- \_\_\_\_\_. Non-Gaussian Multivariate Statistics. <http://homepage.mac.com/shoheishimizu/> Accessed July 9, 2010.
- SPIRITES, P., MEEK C., AND RICHARDSON T. “An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias.” In C. Glymour and G.F. Cooper, eds., *Computation, Causation and Discovery*. Menlo Park: American Association for Artificial Intelligence Press and Cambridge, MA: MIT Press, 1999.
- \_\_\_\_\_, GLYMOUR, C., AND SCHEINES, R. *Causation, Prediction, and Search*. Cambridge, MA: The MIT Press, 2<sup>st</sup> edition, 2001.
- STONE, J. V. *Independent Component Analysis: A Tutorial Introduction*. Cambridge, MA: The MIT Press, 2004.
- YUAN, K.H., AND BENTLER, P.M. “Mean and Covariance Structure Analysis: Theoretical and Practical Improvements.” *Journal of the American Statistical Association*, Vol. 92 (1997), pp. 767-774.
- ZHANG, J. “Causal Reasoning with Ancestral Graphs.” *Journal of Machine Learning Research*, Vol. 9 (2008a), pp. 1437-1474.
- \_\_\_\_\_. “On the Completeness of Orientation Rules for Causal discovery in the Presence of Latent Confounders and Selection Bias.” *Artificial Intelligence*, Vol. 172 (2008b), pp. 1873-1896.
- ZHANG, K., AND HYVÄRINEN, A. “Causality Discovery with Additive Disturbances: An Information-Theoretical Perspective.” *In Proceedings of European Conference on Machine Learning (ECML2009)*, Bled, Slovenia, pp. 570-585, 2009.

## APPENDIX A

### PROPERTIES OF ENTROPY AND NEGENTROPY

Most text of this section is summarized from Chapter 2 and Chapter 8 from Cover and Thomas (2006); Hyvärinen, Karhunen and Oja (2001).

#### Entropy

Suppose two random variables  $X$  and  $Y$  with a joint density function  $f(x, y)$  and marginal probability density function  $f(x)$  and  $f(y)$  respectively.

Then the joint differential entropy  $H(X, Y)$  is defined as

$$H(X, Y) = -\int f(x, y) \log f(x, y) dx dy = -E[\log f(x, y)]$$

The conditional entropy of  $X$  given  $Y$  is the expected value of the entropies of  $X$  conditional on all ranges of  $Y$ :

$$\begin{aligned} H(X|Y) &= -\int f(y) H(X|Y=y) dy = -\int f(y) \left( \int f(x|y) \log f(x|y) dx \right) dy \\ &= -\int f(y) \left( \int \frac{f(x, y)}{f(y)} \log f(x|y) dx \right) dy \\ &= -\int f(y) \frac{1}{f(y)} \left( \int f(x, y) \log f(x|y) dx \right) dy \\ &= -\int f(x, y) \log f(x|y) dx dy \end{aligned}$$

The conditional entropy can be interpreted to be the amount of unknown information of  $X$  after  $Y$  is revealed. Noticeably, when  $X$  and  $Y$  are independent, the conditional entropy of  $X$  given  $Y$  equals to the entropy of  $X$  itself:

$$\begin{aligned}
H(X|Y) &= -\int f(y)H(X|Y=y)dy = -\int f(y)\left(\int f(x|y)\log f(x|y)dx\right)dy \\
&= -\int \int f(x,y)\log f(x)dx dy \\
&= -\int \left(\int f(x,y)dy\right)\log f(x)dx \\
&= -\int f(x)\log f(x)dx = H(X)
\end{aligned}$$

This equation describes that  $Y$  cannot offer any information for the understanding of  $X$  given the independent relationship.

In general,  $f(x|y) = \frac{f(x,y)}{f(y)}$ , so the conditional entropy can be re-written as:

$$\begin{aligned}
H(X|Y) &= -\int f(x,y)\log f(x|y)dx dy = -\int f(x,y)\log \frac{f(x,y)}{f(y)}dx dy \\
&= -\int f(x,y)\log f(x,y)dx dy + \int \log f(y)\left(\int f(x,y)dx\right)dy \\
&= -\int f(x,y)\log f(x,y)dx dy + \int f(y)\log f(y)dy \\
&= H(X,Y) - H(Y)
\end{aligned}$$

Thus,  $H(X,Y) = H(X|Y) + H(Y)$ . Inductively, if  $H(X,Y)$  is necessary to describe and  $X$  and  $Y$  completely when  $H(Y)$  is known, the system still needs the amount of information  $H(X|Y)$  bits. If  $X$  and  $Y$  are independent,  $H(X,Y) = H(X|Y) + H(Y) = H(X) + H(Y)$ . Under this situation, the joint differential entropy is the summation of their individual entropies.

Suppose a single random variable  $X$  is multiplied by a scalar constant,  $w$ . That is

$$y = g(X) = wx \text{ and } x = g^{-1}(y) = \frac{y}{w} \text{ where } g^{-1}(y) \text{ is the inverse function of } g(X).$$

According to the theorem of calculus on the change of variable,  $f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y))$ ,

and

$$\begin{aligned}
 H(y) &= H(wx) = -\int f_Y(y) \log f_Y(y) dy \\
 &= -\int \frac{1}{|w|} f_X\left(\frac{y}{w}\right) \log \left[ \frac{1}{|w|} f_X\left(\frac{y}{w}\right) \right] dy \\
 &= -\int \frac{1}{|w|} f_X\left(\frac{y}{w}\right) \log \frac{1}{|w|} (w dx) - \int \frac{1}{|w|} f_X\left(\frac{y}{w}\right) \log f_X\left(\frac{y}{w}\right) (w dx) \\
 &= -\log \frac{1}{|w|} \int f_X\left(\frac{y}{w}\right) dx - \int f_X\left(\frac{y}{w}\right) \log f_X\left(\frac{y}{w}\right) dx \\
 &= \log |w| + H(x)
 \end{aligned}$$

Suppose  $Y = G(X)$  of  $n$  equations so  $X = G^{-1}(Y)$ ,  $Y$  is written as

$$Y = \begin{bmatrix} G_1(X) \\ G_2(X) \\ \vdots \\ G_n(X) \end{bmatrix} = \begin{bmatrix} G_1(x_1, x_2, \dots, x_n) \\ G_2(x_1, x_2, \dots, x_n) \\ \vdots \\ G_n(x_1, x_2, \dots, x_n) \end{bmatrix}$$

and the Jacobian matrix of the coordinate transformation  $Y = G(X)$  is

$$JB(x_1, \dots, x_n) = \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \dots & \frac{\partial G_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial G_n}{\partial x_1} & \dots & \frac{\partial G_n}{\partial x_n} \end{bmatrix}$$

Taking the differential of  $Y$

$$dY = Y_X dX = |\det(JB(x_1, \dots, x_n))| dX$$

and

$$f_{y_1, \dots, y_n}(y_1, \dots, y_n) = |\det(JB(x_1, \dots, x_n))|^{-1} f_{x_1, \dots, x_n}(G^{-1}(y_1, \dots, y_n))$$

If  $Y = G(X) = WX$ , then  $JB(x_1, \dots, x_n) = W$  and

$$\begin{aligned} H(Y) &= H(WX) = -\int f_Y(y_1, \dots, y_n) \log[f_Y(y_1, \dots, y_n)] dY \\ &= -\int |\det(JB(x_1, \dots, x_n))|^{-1} f_X(G^{-1}(Y)) \log\left(|\det(JB(x_1, \dots, x_n))|^{-1} f_X(G^{-1}(Y))\right) dY \\ &= -\int |\det(W)|^{-1} f_X(G^{-1}(Y)) \log\left(|\det(W)|^{-1}\right) |\det(W)| dX \\ &= -\int |\det(W)|^{-1} f_X(G^{-1}(Y)) \log\left(|\det(W)|^{-1}\right) |\det(W)| dX \\ &\quad - \int |\det(W)|^{-1} f_X(G^{-1}(Y)) \log f_X(G^{-1}(Y)) |\det(W)| dX \\ &= -\log\left(|\det(W)|^{-1}\right) \int f_X(G^{-1}(Y)) dX - \int f_X(G^{-1}(Y)) \log f_X(G^{-1}(Y)) dX \\ &= \log|\det(W)| + H(X) \end{aligned}$$

where  $\det(W)$  denotes the determinant of matrix  $W$ .

### Negentropy

Negentropy  $J$  is defined as

$$J(X) = J(x_1, \dots, x_n) = H(X_{gauss}) - H(X)$$

where  $X_{gauss}$  is a random vector of multivariate Gaussian distribution (multivariate normal distribution) with mean vector  $\mu$  and covariance matrix  $\Sigma$ . The probability density function of  $X_{gauss} \in R^n$  is defined:

$$f(X_{gauss}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det(\Sigma)|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}$$

Obviously, given the information of mean vector (the first-order moment structure) and covariance matrix (the second-order moment structure) is enough to define the multivariate Gaussian probability density completely. It infers that the higher-order

(more than two) moment structures are not required for the understanding of the multivariate Gaussian distribution. Additionally,

$$\begin{aligned}
H(X_{gauss}) &= -\int f(X_{gauss}) \log f(X_{gauss}) dX \\
&= -\int f(X_{gauss}) \left[ \log \left( \frac{1}{(2\pi)^{\frac{n}{2}} |\det(\Sigma)|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)} \right) \right] dX \\
&= -\int f(X_{gauss}) \left[ -\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu) \log e - \log \left[ (2\pi)^{\frac{n}{2}} |\det(\Sigma)|^{\frac{1}{2}} \right] \right] dX \\
&= \frac{1}{2} \log e \cdot \int f(X_{gauss}) [(X-\mu)^T \Sigma^{-1}(X-\mu)] dX + \log \left[ (2\pi)^{\frac{n}{2}} |\det(\Sigma)|^{\frac{1}{2}} \right] \cdot \int f(X_{gauss}) dX \\
&= \frac{1}{2} \log e \cdot E \left[ \sum_{i,j} (x_i - \mu_i) (\Sigma^{-1})_{ij} (x_j - \mu_j) \right] + \frac{1}{2} \log [(2\pi)^n |\det(\Sigma)|] \\
&= \frac{1}{2} \log e \left[ \sum_j \sum_i \Sigma_{ji} \cdot (\Sigma^{-1})_{ij} \right] + \frac{1}{2} \log [(2\pi)^n |\det(\Sigma)|] \\
&= \frac{1}{2} \log e \left[ \sum_{i,j} (\Sigma \Sigma^{-1})_{ij} \right] + \frac{1}{2} \log [(2\pi)^n |\det(\Sigma)|] \\
&= \frac{1}{2} \log e \left[ \sum_{i,j} I_{ij} \right] + \frac{1}{2} \log [(2\pi)^n |\det(\Sigma)|] \\
&= \frac{n}{2} \log e + \frac{1}{2} [\log(2\pi)^n + \log |\det(\Sigma)|] \\
&= \frac{1}{2} \log |\det(\Sigma)| + \frac{n}{2} [\log e + \log(2\pi)]
\end{aligned}$$

Noticeably, with an identical covariance matrix, multivariate normal probability distribution maximizes the entropy over all the distributions. The reason is that:

Let the random vector  $X \in R^n$  have zero mean. Suppose  $g(X)$  is an arbitrary probability density function other than  $f(X_{gauss})$ . Another assumption is  $g(X)$  has the same covariance matrix as  $f(X_{gauss})$ , so  $\int g(X) x_i x_j dX = \Sigma_{ij} = \int f(X_{gauss}) x_i x_j dX$  for all  $i, j$ .

Then, if there is a quadratic form such as  $\log f(X_{gauss})$  exists, then

$\int g(X) \log f(X_{gauss}) = \int f(X_{gauss}) \log f(X_{gauss})$ . Based on the probability density form of multivariate normal probability distribution  $f(X_{gauss})$  and the result of (58), we have

$$\begin{aligned}
0 &\leq D(g(X) \| f(X_{gauss})) = \int g(X) \log \frac{g(X)}{f(X_{gauss})} dX \\
&= \int g(X) \log g(X) dX - \int g(X) \log f(X_{gauss}) dX \\
&= -H(g(x_1, \dots, x_n)) - \int g(X) \log f(X_{gauss}) dX \\
&= -H(g(x_1, \dots, x_n)) - \int f(X_{gauss}) \log f(X_{gauss}) dX \\
&= -H(g(X)) + H(X_{gauss})
\end{aligned}$$

It infers that  $H(g(X)) \leq H(X_{gauss})$ , and the Gaussian distribution maximizes the entropy over all distributions with the same covariance matrix. Because of the above reason, negentropy is always nonnegative and is zero if and only if  $X$  has a Gaussian distribution.

Besides, if there is a linear transformation,  $Y = WX$ , and then negentropy of  $Y$  is

$$\begin{aligned}
J(Y) &= J(WX) = H((WX)_{gauss}) - H(WX) \\
&= \frac{1}{2} \log |\det(W\Sigma W^T)| + \frac{n}{2} [\log e + \log(2\pi)] - [\log |\det(W)| + H(X)] \\
&= \frac{1}{2} \log |\det(W) \det(\Sigma) \det(W^T)| + \frac{n}{2} [\log e + \log(2\pi)] - \log |\det(W)| - H(X) \\
&= \frac{1}{2} \log |\det(\Sigma)| + \log |\det(W)| + \frac{n}{2} [\log e + \log(2\pi)] - \log |\det(W)| - H(X) \\
&= \frac{1}{2} \log |\det(\Sigma)| + \frac{n}{2} [\log e + \log(2\pi)] - H(X) \\
&= \frac{1}{2} \log |\det(\Sigma)| + \frac{n}{2} [\log e + \log(2\pi)] - H(X) \\
&= H(X_{gauss}) - H(X)
\end{aligned}$$



The above formula proves negentropy is scale-invariant for invertible linear transformation. Because of the above properties, negentropy is used to be a non-normality measure in independent component analysis (ICA).

## APPENDIX B

### NOTES OF LINEAR AND MATRIX ALGEBRA

Parts of this appendix's text are summarized from Searle and Willett (2001); Leon (2006). Some components of what follows particularly the statements of matrix operation come directly from the text.

Linear Transformation:

- (1) Let  $L : V \rightarrow W$  be a linear transformation. The kernel of  $L$ , denoted by  $\ker(L)$ , is defined by:

$$\ker(L) : \{v \in V | L(v) = 0_w\}$$

Suppose  $A$  and  $B$  are square, non-singular  $n \times n$  matrices:

- (1) The transpose of a product matrix is the product of the transposed matrices taken in reverse order:  $(AB)^T = B^T A^T$ .
- (2) The inverse of a product matrix is the product of the inverse matrices taken in reverse sequence:  $(AB)^{-1} = B^{-1} A^{-1}$ .
- (3) Suppose  $A$  is an  $n \times n$  symmetric matrix, and then  $\text{vech}(A)$  is to stack the elements of each column of  $A$  which is on and below the main diagonal:

$$vech(A) = \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \\ a_{22} \\ \vdots \\ a_{n2} \\ \vdots \\ a_{(n-1)n} \\ a_{nn} \end{bmatrix}$$

$vech(A)$  removes the redundant elements of  $A$ .

Suppose  $G$  is a  $p \times q$  matrix and  $H$  is an  $m \times n$ , then the Kronecker product of matrix  $G$  and  $H$  is defined as:

$$G \otimes H = \begin{bmatrix} g_{11} & \cdots & g_{1q} \\ \vdots & \ddots & \vdots \\ g_{p1} & \cdots & g_{pq} \end{bmatrix} \otimes H = \begin{bmatrix} g_{11}H & \cdots & g_{1q}H \\ \vdots & \ddots & \vdots \\ g_{p1}H & \cdots & g_{pq}H \end{bmatrix}$$

which is a  $pm \times qn$  matrix.

Suppose  $D$  is an  $n \times n$  diagonal matrix:

(1) The transpose of  $D$  is equal to itself:  $D^T = D$

(2) If there are non-negative entries in  $D$ 's main diagonal, then the square root of  $D$  is calculated by taking the square roots of the main diagonal elements:

$$D^{\frac{1}{2}} = \begin{bmatrix} \sqrt{d_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}, \text{ and thus } D \text{'s decomposition can be written as:}$$

$$D = D^{\frac{1}{2}} D^{\frac{1}{2}}$$

(3) The inverse of  $D$  is calculated by taking the inverses of the main diagonal elements:

$$D^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 & \cdots & 0 \\ 0 & d_{22}^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & d_{nn}^{-1} \end{bmatrix}$$

Properties of Determinants:

(1) The determinant of the transpose of a given square matrix is the same as the

determinant of the matrix itself:  $\det(A^T) = \det(A)$ .

(2) When  $A$  and  $B$  are  $n \times n$  matrices, then the determinant of a matrix product is the

product of the determinants:  $\det(AB) = \det(A)\det(B)$ .

An  $n \times n$  matrix  $Q$  is called an orthogonal matrix, and then

(1)  $Q^T Q = I$ .

(2)  $Q^T = Q^{-1}$ .

(3)  $\|QX\|^2 = \|X\|^2$ .

(4)  $1 = \det(I) = \det(Q^T Q) = \det(Q^T) \det(Q) = [\det(Q)]^2$ , so  $|\det(Q^T)| = |\det(Q)| = 1$

Orthogonal Diagonalizable:

(1)  $A$  is orthogonal diagonalizable if and only if  $A$  is symmetric.

(2)  $A$  is orthogonal diagonalizable if there is an orthogonal matrix  $Q$  such that

$$D = Q^{-1}AQ \text{ is diagonal.}$$

(3) If  $A$  is orthogonal diagonalizable,  $Q$  consists of an orthonormal basis for each

$$\text{eigenspace } E_\lambda = \ker(A - \lambda I).$$

## APPENDIX C

### PROOF OF CONDITONAL DISTRIBUTION OF MULTIVARIATE NORMAL DISTRIBUTION

Most text is summarized from Baba, Shibata, and Sibuya's paper (2004) and *Appendix A: Conditional and Marginal of Multivariate Gaussian*. (n.d.). Retrieved November 14, 2006, from <http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node5.html>.

Assume a two-dimensional random vector,  $V = \begin{bmatrix} Y \\ x_1 \end{bmatrix}$  has a normal distribution

$N(V, \mu, \Sigma)$  with

$$\mu = \begin{bmatrix} E(Y) \\ E(x_1) \end{bmatrix} = \begin{bmatrix} \mu_Y \\ \mu_{x_1} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_{YY} & \sigma_{Yx_1} \\ \sigma_{x_1Y} & \sigma_{x_1x_1} \end{bmatrix}$$

where  $\sigma_{Yx_1} = E((Y - E[Y])(x_1 - E[x_1]))$  and  $\Sigma$  is a symmetric matrix.

The conditional covariance of  $x_1$  and  $x_2$  given  $Y$  is defined:

$$\sigma_{x_1x_2|Y} = Cov(x_1, x_2|Y) = E((x_1 - E(x_1|Y))(x_2 - E(x_2|Y))|Y)$$

Then the conditional covariance matrix is denoted by:

$$\Sigma_{x_1x_2|Y} = \begin{bmatrix} \sigma_{x_1x_1|Y} & \sigma_{x_1x_2|Y} \\ \sigma_{x_2x_1|Y} & \sigma_{x_2x_2|Y} \end{bmatrix}$$

and for the conditional correlation

$$\rho_{x_1x_2|Y} = \frac{\sigma_{x_1x_2|Y}}{\sqrt{\sigma_{x_1x_1|Y}\sigma_{x_2x_2|Y}}}$$

The joint probability density of  $Y$  and  $x_1$  is

$$\begin{aligned} f(V) = f(Y, x_1) &= \frac{1}{(2\pi)\det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(V - \mu)^T \Sigma^{-1}(V - \mu)\right) \\ &= \frac{1}{(2\pi)\det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}Q(Y, x_1)\right) \end{aligned}$$

where

$$\begin{aligned} Q(Y, x_1) &= (V - \mu)^T \Sigma^{-1}(V - \mu) = (Y - \mu_Y, x_1 - \mu_{x_1}) A \begin{pmatrix} Y - \mu_Y \\ x_1 - \mu_{x_1} \end{pmatrix} \\ &= (Y - \mu_Y, x_1 - \mu_{x_1}) \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{pmatrix} Y - \mu_Y \\ x_1 - \mu_{x_1} \end{pmatrix} \\ &= (Y - \mu_Y) a_{11} (Y - \mu_Y) + 2(x_1 - \mu_{x_1}) a_{12} (Y - \mu_Y) + (x_1 - \mu_{x_1}) a_{22} (x_1 - \mu_{x_1}) \end{aligned}$$

where  $A = \Sigma^{-1}$  is symmetric,  $a_{12} = a_{21}$ . Also,  $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \Sigma^{-1}$  and therefore

$$\begin{aligned} a_{11} &= \frac{\sigma_{x_1 x_1}}{\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2} = \frac{1}{\sigma_{YY}} \left( \frac{\sigma_{x_1 x_1} \sigma_{YY}}{\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2} \right) \\ &= \frac{1}{\sigma_{YY}} + \frac{1}{\sigma_{YY}} \left( \frac{\sigma_{x_1 x_1} \sigma_{YY} - [\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2]}{\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2} \right) = \frac{1}{\sigma_{YY}} + \frac{1}{\sigma_{YY}} \left( \frac{\frac{(\sigma_{x_1 Y})^2}{\sigma_{YY}}}{\sigma_{x_1 x_1} - \frac{(\sigma_{x_1 Y})^2}{\sigma_{YY}}} \right) \\ a_{22} &= \frac{\sigma_{YY}}{\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2} = \frac{1}{\sigma_{x_1 x_1} - \frac{(\sigma_{x_1 Y})^2}{\sigma_{YY}}} \\ a_{12} &= \frac{-\sigma_{x_1 Y}}{\sigma_{x_1 x_1} \sigma_{YY} - (\sigma_{x_1 Y})^2} = \frac{-\sigma_{x_1 Y}}{\sigma_{YY}} \frac{1}{\left( \sigma_{x_1 x_1} - \frac{(\sigma_{x_1 Y})^2}{\sigma_{YY}} \right)} = a_{21} \end{aligned}$$

It is defined that

$$G \equiv \sigma_{x_1x_1} - \frac{(\sigma_{x_1Y})^2}{\sigma_{YY}} \text{ and } F \equiv \mu_{x_1} + \frac{\sigma_{x_1Y}}{\sigma_{YY}}(Y - \mu_Y)$$

Hence, the elements of  $A$  can be simplified to

$$a_{11} = \frac{1}{\sigma_{YY}} + \frac{1}{\sigma_{YY}} \left( \frac{\frac{(\sigma_{x_1Y})^2}{\sigma_{YY}}}{\sigma_{x_1x_1} - \frac{(\sigma_{x_1Y})^2}{\sigma_{YY}}} \right) = \frac{1}{\sigma_{YY}} + \frac{1}{\sigma_{YY}} \left( \frac{\frac{(\sigma_{x_1Y})^2}{\sigma_{YY}}}{G} \right)$$

$$a_{22} = \frac{1}{G}$$

$$a_{12} = \frac{-\sigma_{x_1Y}}{\sigma_{YY}} \frac{1}{\left( \sigma_{x_1x_1} - \frac{(\sigma_{x_1Y})^2}{\sigma_{YY}} \right)} = \frac{-\sigma_{x_1Y}}{\sigma_{YY}} \frac{1}{G}$$

As a result,

$$\begin{aligned} Q(x_1, Y) &= (Y - \mu_Y)a_{11}(Y - \mu_Y) + 2(Y - \mu_Y)a_{12}(x_1 - \mu_{x_1}) + (x_1 - \mu_{x_1})a_{22}(x_1 - \mu_{x_1}) \\ &= (Y - \mu_Y) \left( \frac{1}{\sigma_{YY}} + \frac{1}{\sigma_{YY}} \left( \frac{(\sigma_{x_1Y})^2}{G} \right) \right) (Y - \mu_Y) - 2(Y - \mu_Y) \frac{\sigma_{x_1Y}}{\sigma_{YY}} \frac{1}{G} (x_1 - \mu_{x_1}) + \frac{1}{G} (x_1 - \mu_{x_1})^2 \\ &= \frac{1}{\sigma_{YY}} (Y - \mu_Y)^2 + \frac{1}{G} \left( (x_1 - \mu_{x_1})^2 - 2(x_1 - \mu_{x_1}) \frac{\sigma_{x_1Y}}{\sigma_{YY}} (Y - \mu_Y) + \left( \frac{\sigma_{x_1Y}}{\sigma_{YY}} (Y - \mu_Y) \right)^2 \right) \\ &= \frac{1}{\sigma_{YY}} (Y - \mu_Y)^2 + \frac{1}{G} \left( (x_1 - \mu_{x_1}) - \frac{\sigma_{x_1Y}}{\sigma_{YY}} (Y - \mu_Y) \right)^2 \\ &= \frac{1}{\sigma_{YY}} (Y - \mu_Y)^2 + \frac{1}{G} (x_1 - F)^2 \end{aligned}$$

Suppose

$$Q_1(Y) = \frac{1}{\sigma_{YY}} (Y - \mu_Y)^2$$



$$Q_2(Y, x_1) = \frac{1}{G} \left( (x_1 - \mu_{x_1}) - \frac{\sigma_{x_1 Y}}{\sigma_{YY}} (Y - \mu_Y) \right)^2 = \frac{1}{G} (x_1 - F)^2 \quad \text{then}$$

$$Q(Y, x_1) = Q_1(Y) + Q_2(Y, x_1)$$

Then the joint probability density can be written as:

$$\begin{aligned} f(V) = f(Y, x_1) &= \frac{1}{(2\pi) |\det(\Sigma)|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} Q(Y, x_1)\right) \\ &= \frac{1}{(2\pi) |\sigma_{YY}|^{\frac{1}{2}} \left| \sigma_{x_1 x_1} - \frac{(\sigma_{x_1 Y})^2}{\sigma_{YY}} \right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} Q(Y, x_1)\right) \\ &= \frac{1}{(2\pi)^{\frac{1}{2}} |\sigma_{YY}|^{\frac{1}{2}}} \exp\left(-\frac{(Y - \mu_Y)^2}{2\sigma_{YY}}\right) \frac{1}{(2\pi) |G|^{\frac{1}{2}}} \exp\left(-\frac{1}{2G} (x_1 - F)^2\right) \\ &= N(Y, \mu_Y, \sigma_{YY}) N(x_1, F, G) = f(Y) N(x_1, F, G) \end{aligned}$$

Therefore, the conditional normal distribution of  $x_1$  given  $Y$  is

$$f(x_1|Y) = \frac{f(Y, x_1)}{f(Y)} = \frac{1}{(2\pi) |G|^{\frac{1}{2}}} \exp\left(-\frac{1}{2G} (x_1 - F)^2\right)$$

with mean

$$E(x_1|Y) = F = \mu_{x_1} + \frac{\sigma_{x_1 Y}}{\sigma_{YY}} (Y - \mu_Y)$$

We can calculate  $E(x_2|Y)$  following the same procedure and then it is easy to derive the conditional covariance of  $x_1$  and  $x_2$  given  $Y$ .

## APPENDIX D

### PROOF OF EXACT FORM ON $J = \frac{\partial \sigma_2(\hat{\tau})}{\partial \hat{\tau}^T}$

Most text here is summarized from the Appendix E of Shimizu, Hoyer, Hyvärinen, and Kerminen's paper (2006).

Suppose  $n$  is the number of the dimension of  $X$ . The model-based variance-covariance matrix of  $X$ ,  $\Sigma$ , is defined as:

$$\Sigma = E(XX^T) = \left[ D^{-\frac{1}{2}}(I - B) \right]^{-1} \left\{ \left[ D^{-\frac{1}{2}}(I - B) \right]^{-1} \right\}^T = YY^T$$

where  $Y = \left[ D^{-\frac{1}{2}}(I - B) \right]^{-1}$  is a lower triangular matrix,  $B$  is a strictly lower triangular matrix, and  $D$  is a diagonal matrix. Obviously,  $\Sigma$  is a function of  $Y$ , which is in turn a function of  $B$  and  $D$ . Therefore,

$$\begin{aligned} \frac{\partial \Sigma_{ij}}{\partial b_{kl}} &= \frac{\partial (YY^T)_{ij}}{\partial b_{kl}} = \sum_p \sum_q \frac{\partial (YY^T)_{ij}}{\partial Y_{pq}} \frac{\partial Y_{pq}}{\partial b_{kl}} \\ \frac{\partial \Sigma_{ij}}{\partial d_{kk}} &= \frac{\partial (YY^T)_{ij}}{\partial d_{kk}} = \sum_p \sum_q \frac{\partial (YY^T)_{ij}}{\partial Y_{pq}} \frac{\partial Y_{pq}}{\partial d_{kk}} \end{aligned}$$

where  $b_{kl}$  and  $d_{kk}$  are scalar from elements of  $B$  and  $D$ , respectively, and

$$\frac{\partial (YY^T)_{ij}}{\partial Y_{pq}} = \begin{cases} 2Y_{pq} & (i = p, j = p) \\ Y_{jq} & (i = p, j \neq p) \\ Y_{iq} & (i \neq p, j = p) \\ 0 & (i \neq p, j \neq p) \end{cases}$$

Besides, since  $YY^{-1} = I$

$$\frac{\partial(YY^{-1})}{\partial b_{kl}} = \frac{\partial Y}{\partial b_{kl}} Y^{-1} + Y \frac{\partial Y^{-1}}{\partial b_{kl}} = 0$$

$$\frac{\partial(YY^{-1})}{\partial d_{kk}} = \frac{\partial Y}{\partial d_{kk}} Y^{-1} + Y \frac{\partial Y^{-1}}{\partial d_{kk}} = 0$$

from which

$$\frac{\partial Y}{\partial b_{kl}} = -Y \frac{\partial Y^{-1}}{\partial b_{kl}} Y$$

$$\frac{\partial Y}{\partial d_{kk}} = -Y \frac{\partial Y^{-1}}{\partial d_{kk}} Y$$

where  $Y^{-1} = D^{-\frac{1}{2}}(I - B)$ . Thus,

$$\frac{\partial Y^{-1}}{\partial b_{kl}} = -D^{-\frac{1}{2}} L^{kl}$$

$$\frac{\partial Y^{-1}}{\partial d_{kk}} = -\frac{1}{2} (d_{kk})^{-\frac{3}{2}} L^{kk} (I - B)$$

where  $L^{kl}$  is a  $n \times n$  matrix with 1 at  $k$ th row and  $l$ th column, and zero otherwise.

Finally, we have

$$\frac{\partial Y}{\partial b_{kl}} = Y D^{-\frac{1}{2}} L^{kl} Y$$

$$\frac{\partial Y}{\partial d_{kk}} = Y \frac{(d_{kk})^{-\frac{3}{2}}}{2} L^{kk} (I - B) Y$$

As a result, in our case, when  $n = 3$

$$\begin{aligned}
J &= \frac{\partial \sigma_2(\hat{\tau})}{\partial \hat{\tau}^T} = \frac{\partial \text{vech}(\Sigma)}{\partial \hat{\tau}^T} \\
&= \left[ \begin{array}{ccccc} \frac{\partial \text{vech}(\Sigma)}{\partial b_{21}} & \frac{\partial \text{vech}(\Sigma)}{\partial b_{32}} & \frac{\partial \text{vech}(\Sigma)}{\partial d_{11}} & \frac{\partial \text{vech}(\Sigma)}{\partial d_{22}} & \frac{\partial \text{vech}(\Sigma)}{\partial d_{33}} \end{array} \right] \\
&= \left[ \begin{array}{ccccccc} \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{11}}{\partial b_{21}} & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{11}}{\partial b_{32}} & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{11}}{\partial d_{11}} & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{11}}{\partial d_{33}} \\ \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{21}}{\partial b_{21}} & \dots & \dots & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{21}}{\partial d_{33}} \\ \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{31}}{\partial b_{21}} & \dots & \dots & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{31}}{\partial d_{33}} \\ \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{22}}{\partial b_{21}} & \dots & \dots & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{22}}{\partial d_{33}} \\ \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{32}}{\partial b_{21}} & \dots & \dots & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{32}}{\partial d_{33}} \\ \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{33}}{\partial b_{21}} & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{33}}{\partial b_{32}} & \dots & \dots & \frac{\partial(\mathbf{Y}\mathbf{Y}^T)_{33}}{\partial d_{33}} \\ \frac{\partial b_{21}}{\partial b_{21}} & \frac{\partial b_{32}}{\partial b_{32}} & \dots & \dots & \frac{\partial d_{33}}{\partial d_{33}} \end{array} \right]
\end{aligned}$$

## APPENDIX E

The text offers the proof of that the third- and fourth-order moment structures of Model 1' :  $y = \beta x + \varepsilon_y$  and Model 2' :  $x = \eta y + \varepsilon_x$  are different to each other. Most proof here is summarized from Shimizu and Hoyer's paper (2008).

First, following the rule of ordinary least square method,  $\beta$  and  $\eta$  can be written as the functions of covariance between  $x$  and  $y$  and variances of  $x$  and  $y$  like:

$$\beta = \frac{\text{cov}(x, y)}{\text{var}(x)}, \quad \eta = \frac{\text{cov}(x, y)}{\text{var}(y)}.$$

Hence, the square of correlation coefficient is equal to  $\beta\eta$  and it is required here that

$$0 < \beta\eta < 1$$

Suppose  $x$  and  $y$  are centered and have finite fourth-order moments. The expected values of error terms are assumed to be zero. Let  $\sigma_i^{(1)}(\hat{\tau}_i^{(1)})$  and  $\sigma_i^{(2)}(\hat{\tau}_i^{(2)})$  denote the  $i$ -th moment structures of Model 1' and 2', respectively. Also, fourth cumulant of a random variable  $z$  with  $E(z) = 0$  is denoted by  $\kappa_4(z) = E(z^4) - 3E(z^2)^2$ .

Denote  $m_{ij} = \frac{1}{N} \sum_{k=1}^N x_k^i y_k^j$ . When Model 1' holds true, then its third-order moments

are that:

$$E \begin{bmatrix} m_{30} \\ m_{21} \\ m_{12} \\ m_{03} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \\ \beta^2 & 0 \\ \beta^3 & 1 \end{bmatrix} \begin{bmatrix} E(x^3) \\ E(\varepsilon_y^3) \end{bmatrix} = \sigma_3^{(1)}(\hat{\tau}_3^{(1)}) \quad (\text{E1})$$

On the other hand, if Model 2' holds true, then the third-order moments of Model 2' are:

$$E \begin{bmatrix} m_{30} \\ m_{21} \\ m_{12} \\ m_{03} \end{bmatrix} = \begin{bmatrix} \eta^3 & 1 \\ \eta^2 & 0 \\ \eta & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} E(y^3) \\ E(\varepsilon_x^3) \end{bmatrix} = \sigma_3^{(2)}(\hat{\tau}_3^{(2)}) \quad (\text{E2})$$

Once we assume the third-order moments of Model 1' and Model 2' are equivalent, then the following conditions hold:

$$\beta E(x^3) = \eta^2 E(y^3) \quad \text{and} \quad \beta^2 E(x^3) = \eta E(y^3)$$

from which it follows that

$$\beta(1 - \beta\eta)E(x^3) = 0$$

Since  $\beta\eta < 1$  has been assumed, then the condition, satisfying that the third order moment structures of Model 1' and Model 2' are the same, is  $E(x^3) = 0$ . However,  $x$  is required to be non-normal distributed so it is impossible to have  $E(x^3) = 0$ . Furthermore, using (E1) and (E2), we have  $E(x^3) = 0$  and therefore  $E(\varepsilon_y^3) = 0$ . However, because the exogenous variable and error term are assumed to be non-normal,  $E(x^3) = E(\varepsilon_y^3) = 0$  does not hold. Thus, the conclusion is that  $\sigma_3^{(1)}(\hat{\tau}_3^{(1)}) \neq \sigma_3^{(2)}(\hat{\tau}_3^{(2)})$  with non-normal variables and error terms.

Moreover, the fourth-order moments can apply the same framework as above discussion. That is:

$$E \begin{bmatrix} m_{40} \\ m_{31} \\ m_{22} \\ m_{13} \\ m_{04} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \beta & 0 \\ \beta^2 & 0 \\ \beta^3 & 0 \\ \beta^4 & 1 \end{bmatrix} \begin{bmatrix} \kappa_4(x) \\ \kappa_4(\varepsilon_Y) \end{bmatrix} = \sigma_4^{(1)}(\hat{\tau}_4^{(1)}) \quad (\text{E3})$$

and

$$E \begin{bmatrix} m_{40} \\ m_{31} \\ m_{22} \\ m_{13} \\ m_{04} \end{bmatrix} = \begin{bmatrix} \eta^4 & 1 \\ \eta^3 & 0 \\ \eta^2 & 0 \\ \eta & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \kappa_4(y) \\ \kappa_4(\varepsilon_X) \end{bmatrix} = \sigma_4^{(2)}(\hat{\tau}_4^{(2)}) \quad (\text{E4})$$

Similarly, if  $\sigma_4^{(1)}(\hat{\tau}_4^{(1)}) = \sigma_4^{(2)}(\hat{\tau}_4^{(2)})$  is presumed, then it follows  $\beta(1 - \beta\eta)\kappa_4(x) = 0$ . Since the correlation coefficient between  $x$  and  $y$  is set to be not equal to 1 or -1, therefore  $\kappa_4(x) = 0$  and then  $\kappa_4(\varepsilon_Y) = 0$ . Nevertheless,  $\kappa_4(x) \neq 0$  and  $\kappa_4(\varepsilon_Y) \neq 0$  due to non-normal distribution and thus  $\sigma_4^{(1)}(\hat{\tau}_4^{(1)}) = \sigma_4^{(2)}(\hat{\tau}_4^{(2)})$  does not hold.

## APPENDIX F

**TABLE F1 OLS Results of Estimated Coefficient on Distance Metrics of Disaggregate CSD (Package Size is Treated as a Continuous Variable)**

Distance Metrics	Cross-Price	
	Coeff	t-stat.
Continuous Distance Measures with Continuous Variables		
<i>One-Dimensional</i>		
Carbohydrate Content (Mcarb)	-2.68*	-2.85
Sodium Content (Mso)	0.41	0.38
Container Volume (Mvol)	-0.34	-0.90
<i>Two-Dimensional</i>		
Sodium/Carbohydrate Content (MSC)	31.84*	6.45
Sodium/Volume Content (MSV)	-23.98*	-16.05
Carbohydrate/Volume Content (MCV)	19.20*	13.48
Discrete Distance Measures with Continuous Variables		
<i>Nearest Neighbor</i>		
Sodium/Carbohydrate Content (MNNSC)	-26.86*	-5.74
Sodium/Carbohydrate Content with Price (MNNSCP)	7.41*	4.07
Sodium/Carbohydrate/Volume Content (MNNSCV)	-2.09	-1.2
<i>Common Boundaries</i>		
Sodium/Carbohydrate Content (MCBSC)	2.09**	2.54
Carbohydrates/Volume Content (MCBCV)	1.4	1.74
Discrete Distance Measures with Discrete Variables		
<i>Product Classifications</i>		
Manufacturer Identity (Mmanu)	4.35	0.55
Brand Identity (Mbrand)	28.24*	4.08
Product grouping (Mgroup)	-51.07*	-5.93
Citric Acid Containing (Mcitric)	15.53	0.98

All regressions include cluster, product, and year dummy indicators.

Coefficients have been multiplied by 1,000 for readability

\* Significant at 1% , \*\* Significant at 5%



**TABLE F2 OLS Results of Estimated Coefficient on Distance Metrics of Disaggregate CSD (Package Size is Treated as a Dummy Variable)**

<b>Distance Metrics</b>	<b>Cross-Price</b>	
	<b>Coeff</b>	<b>t-stat.</b>
<b>Continuous Distance Measures with Continuous Variables</b>		
<i>One-Dimensional</i>		
Carbohydrate Content (Mcarb)	2.86*	2.82
Sodium Content (Mso)	-5.24*	-5.50
<i>Two-Dimensional</i>		
Sodium/Carbohydrate Content (MSC)	-26.92**	-2.27
<b>Discrete Distance Measures with Continuous Variables</b>		
<i>Nearest Neighbor</i>		
Sodium/Carbohydrate Content (MNNSC)	30.1*	2.77
Sodium/Carbohydrate/Price Content (MNNSCP)	8.8*	5.18
<i>Common Boundaries</i>		
Sodium/Carbohydrates Content (MCBSC)	1.65*	3.14
<b>Discrete Distance Measures with Discrete Variables</b>		
<i>Product Classifications</i>		
Manufacturer Identity (Mmanu)	-14.62	-1.70
Brand Identity (Mbrand)	15.24**	2.18
Cola Product Grouping (Mcola)	12.79	1.21
Caffeine Product Grouping (Mcaff)	-18.86	-1.28
Size classification (Msize)	1.53	0.55
Citric Acid Containing (Mcitric)	24.73	1.70

All regressions include cluster, product, and year dummy indicators.

Coefficients have been multiplied by 1,000 for readability

\* Significant at 1%, \*\* Significant at 5%

**TABLE F3 The Comparison between Our Disaggregate Estimation Results and Dube's Results**

Activities and Attributes	Our result		Dube's result	
	Sign or explanation	Significant	Sign or explanation	Significant
On Promotion	Promotion activity stimulates the consumption.	Yes	Both feature ads and displays have a positive influence on perceived product quality.	yes
Brand	There is a stronger substitution between products with different size of a given brand.	Yes	Consumer has strong loyalty to specific brand than to a given product.	yes
Manufacturer	Consumers do not have tendency to substitute between the products produced by the same manufacturer.	No	unknown	unknown
Package Size	Either	No	positive	yes

Dube's result is summarized from his papers, "Multiple Discreteness and Product Differentiation: Demand for Carbonated Soft Drinks" and "Product Differentiation and Mergers in the Carbonated Soft Drink Industry" published in 2004 and 2005 respectively.

**TABLE F4 List of Aggregate Brands with the Average Retail Price and Their Shares of All Sold Volume (Ordered by Aggregate Sales Volume Shares)**

<b>Product Description</b>	<b>Average Retail Price (\$/fl. oz.)</b>	<b>Cluster A</b>	<b>Cluster B</b>	<b>Cluster C</b>	<b>Cluster D</b>	<b>Total</b>
Pepsi	0.0258	0.0554	0.0327	0.0811	0.0807	0.2499
Coke	0.027	0.0500	0.0258	0.0424	0.0605	0.1787
Diet Pepsi	0.0242	0.0323	0.0116	0.0283	0.0419	0.1140
7 Up	0.0256	0.0240	0.0167	0.0326	0.0281	0.1013
Diet Coke	0.0251	0.0332	0.0104	0.0193	0.0378	0.1007
Diet 7 Up	0.0254	0.0103	0.0042	0.0094	0.0117	0.0356
Diet Caffeine Free Pepsi	0.0257	0.0105	0.0026	0.0069	0.0139	0.0340
Dr Pepper	0.0254	0.0087	0.0030	0.0080	0.0130	0.0328
Sprite	0.0252	0.0086	0.0046	0.0079	0.0102	0.0313
Diet Caffeine Free Coke	0.0267	0.0105	0.0023	0.0040	0.0115	0.0283
Canada Dry Ginger Ale	0.0195	0.0061	0.0044	0.0057	0.0068	0.0231
Mountain Dew Soda	0.0256	0.0057	0.0024	0.0046	0.0087	0.0215
A&W Root Beer	0.0225	0.0034	0.0013	0.0030	0.0045	0.0122

**TABLE F5 Attributes of CSD Brands Used in Our Dataset**

Manufacturer	Product	Calories	Sodium (mg)	Carbohydrates (g)	Caffeine	Contain Citric Acid	Cola
Coca Cola	Coke	140	50	39	1	0	1
	Diet Coke	0	40	0	1	1	1
	Diet Caffeine Free Coke	0	40	0	0	1	1
	Sprite	140	70	38	0	1	0
PepsiCo	Pepsi	150	35	41	1	1	1
	Diet Pepsi	0	35	0	1	1	1
	Diet Caffeine Free Pepsi	0	35	0	0	1	1
	Mountain Dew Soda	170	70	46	1	1	0
Cadbury	Dr Pepper	150	55	40	1	0	0
	7 Up	140	75	39	0	1	0
	Diet 7 Up	0	35	0	0	1	0
	Canada Dry Ginger Ale	140	50	36	0	1	0
	A&W Root Beer	170	65	47	0	0	0

Characteristics are per 12-oz serving.

Data Source: 1. <https://www.wegmans.com/webapp/wcs/stores/servlet/HomepageView?storeId=10052&catalogId=10002&langId=-1>

2. [http://www.pepsiproductfacts.com/infobycategory\\_print.php?pc=p1062&t=1026&s=8&i=ntrtn](http://www.pepsiproductfacts.com/infobycategory_print.php?pc=p1062&t=1026&s=8&i=ntrtn)

**VITA**

Name: Pei-Chun Lai

Address: 2124 TAMU, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843-2124

Email Address: jemmylai.tw@yahoo.com.tw

Education: B.A., Economics, National Taiwan University, 2003  
M.S., Economics, Texas A&M University, 2005  
Ph.D., Agricultural Economics, Texas A&M University, 2010