2015

MODELING LARGE-SCALE CROSS EFFECT IN CO-PURCHASE INCIDENCE: COMPARING ARTIFICIAL NEURAL NETWORK TECHNIQUES AND MULTIVARIATE PROBIT MODELING

Zhiguo Yang
University of Kentucky, yzg1236@hotmail.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Zhiguo Yang, Student

Dr. Devanathan Sudharshan, Major Professor

Dr. Kenneth R. Troske, Director of Graduate Studies

MODELING LARGE-SCALE CROSS EFFECT IN CO-PURCHASE INCIDENCE: COMPARING ARTIFICIAL NEURAL NETWORK TECHNIQUES AND MULTIVARIATE PROBIT MODELING

DISSERTATION

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy in the

College of Business and Economics

at the University of Kentucky

By

Zhiguo Yang

Lexington, Kentucky

Co-Directors: Dr. Devanathan Sudharshan, James and Diane Stuckert BS/MBA Endowed Chair and Professor of Marketing

and : Dr. Clyde Holsapple, Rosenthal Endowed Chair and Professor in Management Information Systems

Lexington, Kentucky

2015

ABSTRACT OF DISSERTATION

MODELING LARGE-SCALE CROSS EFFECT IN CO-PURCHASE INCIDENCE: COMPARING ARTIFICIAL NEURAL NETWORK TECHNIQUES AND MULTIVARIATE PROBIT MODELING

This dissertation examines cross-category effects in consumer purchases from the big data and analytics perspectives. It uses data from Nielsen Consumer Panel and Scanner databases for its investigations. With big data analytics it becomes possible to examine the cross effects of many product categories on each other. The number of categories whose cross effects are studied is called category scale or just scale in this dissertation. The larger the category scale the higher the number of categories whose cross effects are studied. This dissertation extends research on models of cross effects by (1) examining the performance of MVP model across category scale; (2) customizing artificial neural network (ANN) techniques for large-scale cross effect analysis; (3) examining the performance of ANN across scale; and (4) developing a conceptual model of spending habits as a source of cross effect heterogeneity. The results provide researchers and managers new knowledge about using the two techniques in large category scale settings The computational capabilities required by MVP models grow exponentially with scale and thus are more significantly limited by computational capabilities than are ANN models. In our experiments, for scales 4, 8, 16 and 32, using Nielsen data, MVP models could not be estimated using baskets with 16 and more categories. We attempted to and could calibrate ANN models, on the other hand, for both scales 16 and 32. Surprisingly, the predictive results of ANN models exhibit an inverted U relationship with scale. As an ancillary result we provide a method for determining the existence and extent of non-linear own and cross category effects on likelihood of purchase of a category using ANN models. Besides our empirical studies, we draw on the mental budgeting model and impulsive spending literature, to provide a conceptualization of consumer spending habits as a source of heterogeneity in cross effect context. Finally, after a

discussion of conclusions and limitations, the dissertation concludes with a discussion of open questions for future research.

KEYWORDS: Cross category, Co-purchase, Large scale analysis, Multivariate probit model, Artificial neural network.

Zhiguo Yang

Student's signature

11/19/2015

Date

MODELING LARGE-SCALE CROSS EFFECT IN CO-PURCHASE INCIDENCE: COMPARING ARTIFICIAL NEURAL NETWORK TECHNIQUES AND MULTIVARIATE PROBIT MODELING

By

Zhiguo Yang

Dr. Devanathan Sudharshan
Co-Director of Dissertation

Dr. Clyde Holsapple
Co-Director of Dissertation

Dr. Kenneth R. Troske
Director of Graduate Studies

11/19/2015
Date

Dedicated to my parents, Chunxia Yang and Jingfan Yang, who brought me to this world,

and to my wife, Dandan Liu, who is my life partner in this world.

ACKNOWLEDGMENTS

I am forever indebted to my dissertation co-chairs. Without Dr. D Sudharshan's support and guidance, this dissertation development would not be possible. His expertise, sharp and critical-thinking mind, and endless passion in doing research not only inspired me in developing this dissertation, but also will keep inspiring me for the rest of my academic life. Dr. Clyde Holsapple has been mentoring me since day-one of my PhD training. He is always encouraging, patient and caring. His guidance ensured the dissertation development being on the right track. PhD training is a life changing journey for me and working with Dr. D Sudharshan and Dr. Clyde Holsapple is a critical procedure making that change happened.

I would like to thank the dissertation committee and examiner, respectively: Dr. Anita Lee-Post, Dr. Adib Bagh, Dr. Judy Goldsmith, and Dr. Yoonbai Kim. Their feedback, suggestions, and insights helped a lot in improving the quality of this dissertation. I want thank Dr. Goldsmith for providing access to the computing clusters at University of Kentucky's KAOS lab. Paul Eberhart at KAOS lab helped me use those machines, and his timely support allows my data analysis going smoothly.

I am incredibly grateful to my fellow doctoral students and friends. From day to day, we discuss research topics, career concerns, and meanings of life… have fun on weekend days or just simply sit next to each other and edit a term paper. How could I live without your accompany and support in the past 4 years?!

I want to thank my wife, Dandan Liu. Without her scarifies and support in the past 4 years, finishing my PhD study is impossible. I also want to take this opportunity to thank Dr. Jim Baker and Tammi Sutton for their help and support to myself and my family during the past many years.

TABLE of CONTENTS

Acknowledgments……iii
List of Tables……vii
List of Figures……viii

Chapter 1. Introduction 1
1.1. Cross Effect and Large-Scale Cross Effect 1
1.2. Theoretical and Managerial Implications 6
1.3. Related Research 8
1.3.1 Association Rule Mining 8
1.3.2 Affinity Analysis 10
1.4. Summary of Research Question 10
Chapter 2. Background and Literature 11
2.1 The Random Utility Model 11
2.2 Existing Models of Cross Effect 12
2.2.1 Utility Correlation and Multivariate Probit Model (MVP) 12
2.2.2 Conditional Choice and Multivariate Logistic Model 13
2.2.3 Consumption Satiation Model 15
2.2.4 Studies Using an Extended Number of Categories 15
2.3 Modeling Large-Scale Cross Effect 18
2.3.1 Parameter Explosion 18
2.3.2 Reliability of Parameter Estimation 18
2.3.3 Complexity and Effectiveness of Parameters Estimation 19
2.4 Cross Effect Concept Development 21
2.5 Consumer Heterogeneity and Spending habit 23
2.5.1 Heterogeneity in Cross Effect Literature 24
2.5.2 Segmentation Method in Marketing Literature 26
2.5.3 Spending Habits, the Heterogeneity in This Dissertation 28
2.5.4 Data Sparseness, Fixed Effect and Random Effect Model of Heterogeneity 30
2.6 Artificial Neural Network (ANN) Technique 32
2.6.1 Advantages of ANN for Cross Effect Analysis 33
2.6.2 Selected Studies of ANN Application in Marketing Problems 35
2.6.3 Prediction Model 37

Chapter 3. Model Specification and Study Design 38
3.1 MVP Model with Heterogeneous Spending habit 38
3.1.1 The First Level of the MVP Model 39
3.1.2 The second level of the MVP model 42
3.1.3 Data Sparseness and Alternative Solution 44
3.1.4 Bayesian Inference and Monte Carlo Markov Chain (MCMC) Methods 45
3.2 ANN Model 46
3.2.1 The General Construct of ANN 46
3.2.2 Cross Effect ANN model 51
3.2.3 Consumer Heterogeneity in ANN 52
3.2.4 Configuration Tuning for Cross Effect Analysis 53
3.3 Data 57
3.3.1 Four Levels of Category Scale 57
3.3.2 Steps and Conditions Used in Data Extraction 58
3.3.3 Resulting Data Statistics 59
3.4 Study – 1 Large-Scale MVP and Spending Habit Heterogeneity 62
3.4.1 Experiments to Study Model Performance with Increasing Scales 62
3.4.2 Prediction Hit Rate, and Measures Used in This Study 63
3.4.3 Parameter Estimation 68
3.5 Study – 2 Large-Scale ANN and Non-linear Effect 69
3.5.1 Experiments to Study Model Performance with Increasing Scales 69
3.5.2 Nonparametric Model ANN 69
3.5.3 The Special Feature Reported by ANN Model 70
Chapter 4. Results of Study One 72
4.1 Model Runs 72
4.2 Model Performance of Increasing Scale 73
Chapter 5. Results of Study Two 81
5.1 Model Runs 81
5.1.1 Convergence 81
5.2 Model Performance with Increasing Scale 85
5.3 Non-linear Relationship 87
Chapter 6. Discussion 93
6.1 General Comparison between the ANN and the MVP 93
6.2 Findings of the MVP Model Estimation 95
6.3 Data Sparseness and Its Impact on the MVP and ANN Models 96

6.4 Non-linear Relationship, and the General Effect of an ANN Input 97
6.5 Complexity of MVP and ANN models 98
Chapter 7. Conclusions, Limitations and Future Research 99
7.1 Conclusions and Limitations 99
7.2 Future Research 101
7.2.1 Theory of Cross Effect between Unfamiliar Pairs of Categories 101
7.2.2 The Spending Habit Heterogeneity Model and Propositions 103
7.2.3 ANN Incorporating Existing Knowledge 104
7.2.4 Evolutionary Learning Algorithm 105
7.2.5 Impact of Data Preparation on Research Findings and Drawing Conclusions 106
7.3 General Conclusion 106
Appendix 109
Appendix A.1 Representative cross effect (CE) literature 110
Appendix A.2 Pair-wise Joint purchase frequency 112
Appendix A.3 Price parameters draws of the four base category 115
Appendix B Inputs dependency in ANN 122
Appendix C Mean of Percentage Error (MPE) 124
References 127
VITA 134

LIST OF TABLES

Table 2.1 Selected representative literature of cross effect 17
Table 3.1 Variable naming conventions used in model specification 38
Table 3.2 Variable abbreviation conventions used in model specification 38
Table 3.3 Effectiveness of adjusted cross entropy error 55
Table 3.4 Data extraction information 59
Table 3.5 Category List (highlighted are the four base categories) 60
Table 3.6 Part of pair-wise Joint purchase frequency 61
Table 3.7 Example of market basket composition and predictions 63
Table 3.8 Hit rate measure on scale 2 and scale 4 model outcomes 65
Table 3.9 Demonstration of hit rate lift 67
Table 3.10 Calculation of base 4 categories from a scale 8 model 68
Table 4.1 MCMC parameters 72
Table 4.2 General model performance with increasing category scale 74
Table 4.3 Own/Cross Effect estimation (average of 10 runs) 76
Table 4.4 Parameter estimation dispersion of 4 categories and 8 categories model . 79
Table 5.1 Summary of ANN model configurations 81
Table 5.2 Selected error tracking plot 83
Table 5.3 General model performance with increasing category scale [(a)] 85
Table 6.1 General Performance of the ANN and the MVP model 94
Table 6.2 Operational feasibility 94

LIST OF FIGURES

Figure 3.1 Single neuron perceptron, adopted from (West et al. 1997) 47
Figure 3.2 One Hidden Layer NN, adopted from (West et al. 1997) 48
Figure 3.3 Conceptual decision making model 49
Figure 3.4 the cross category ANN model 52
Figure 5.1 Selected plot of generalized weights. 89, 90
Figure 6.1 Number of parameters and computation time 93
Figure 7.1 Conceptual model of spending habit heterogeneity 103

Chapter 1. Introduction

1.1. Cross Effect and Large-Scale Cross Effect

Many consumers supposedly purchase frosting together with cake mix. Research of this and related phenomenon is usually called co-purchase incidence (COPI). In general, a market action, such as price discount or promotion of cake mix increases its sales. Thus, it is reasonable to speculate that the campaign not only increases its own sales, but also increases sales of frosting because many consumers supposedly purchase them together. Quantifying such a "spilled" effect of marketing campaign among categories is related to cross effect analysis in marketing literature (Manchanda et al. 1999). The theoretical and managerial implications of cross effect are well acknowledged such as boosting cross sales (Manchanda et al. 1999), maximizing overall store profit (Wedel and Zhang 2004, Song and Chintagunta 2006, Leeflang and Parreño-Selva 2012, Pancras et al. 2013), and finding new market by identifying cross-selling opportunities (Li et al. 2005).

Most existing literature report models of cross effect with pre-specified "familiar" categories such as bacon and egg, detergent and softener, coffee and tea, and kitchen towel and napkin. The implicit requirement is to have good knowledge about complementarity/substitutability between categories. However, a modern retailer typically carries hundreds, or thousands of categories. Choosing from a list of similar categories for cross effect analysis can be difficult and will constrain research findings within the selected categories. For example, here is a partial list of similar categories in a grocery database:

{ …
FROSTING READY-TO-SPREAD
MIXES - CAKE/SPECIALTY - OVER 10 OZ.
MIXES - CAKE/SPECIALTY - 10 OZ & UNDER
MIXES - CAKE/LAYER - OVER 10 OZ.
MIXES - CAKE/LAYER - 10 OZ & UNDER
FOOD COLORING
EGG COLORING KITS/DYE
MIXES - BROWNIES
MIXES-COFFEE CAKE
MIXES-DESSERT-MISC.
MIXES-COOKIE
MIXES - HUSHPUPPY
CAKE DECORATIONS & ICING
MIXES-FROSTING
MIXES-PIE CRUST
MIXES-DUMPLING & KUGEL
MIXES – PANCAKE
…}.

There are various categories of cake mix and frosting. Making a choice of WHICH mix and WHICH frosting is entered into the cross effect model from this list (which is already truncated) is not an easily justifiable decision. Additionally, pre-specifying a pair of categories excludes the possibility of pairing categories by data-evidence. An ideal model for managers is the one being able to simultaneously model all these categories, i.e. large-scale cross effect model. Large-scale cross effect is the cross effect over a large number of categories in which prior assumptions of possible cross effect is not required. This dissertation examines an existing econometric model (MVP) and the artificial neural network (ANN) technique as candidates for this ideal model. In so doing, it extends the existing literature of cross category research, examines existing models' applications in the big data context, and applies the ANN technique into cross

category research to shed light on alternative approaches of modeling cross category purchase.

The first objective of this dissertation is to examine the performance of existing and alternative models. At the most general level, the research question is: how do the two models perform in large-scale cross effect analysis that simultaneously loads a large number of categories without prior assumptions of complementarity/substitutability? Examples of prior assumptions include knowing that bacon and eggs are commonly used together in a classic American breakfast, knowing that detergent and softener are commonly used together to wash clothes, and knowing that paper towels and napkins may be substitutes. By relaxing this type of prior assumptions, this study allows novel combinations of categories as cross effect partners and bases such relationships on data evidence.

This study examines two models to shed light on the first research question. A multivariate probit model (MVP) from existing literature, and an artificial neural network (ANN) model customized for the cross effect analysis. Using Nielsen's Consumer Panel and Scanner dataset, this study extracts four datasets representing four increasing scales, i.e., four categories, eight categories, 16 categories, and 32 categories. The two models are fit to each of the four datasets. Fitting to same dataset provides a ground for comparing performance. Several other tactics are taken to make a fair comparison. Because ANN allows setting the desirable level of error in model estimation, this study adjusts it to a specific level so that the model's prediction accuracy is similar to that of MVP model (best reachable accuracy is also reported). To reduce the interference of running

environment, the same hardware and operating system settings are used to run the two models.

The results show that, at similar prediction accuracy level and using the same dataset, the ANN model usually finishes computation using much less time and using much fewer computational resources. When scale increases from four to eight and 16 categories, computation time of MVP models are about 84, 850, and (estimated) more than 19,320 minutes (322+ hours), respectively; while ANN models use about 3, 93 and 176 minutes, respectively. The MVP model becomes computationally cumbersome when scale increases to 16 categories, while ANN model can be computed in three hours on average. When scale increases to 32 categories, a PC with 16GB memory hits the out-of-memory error in the middle of computing the MVP model; while the corresponding ANN model can be computed in four hours, on average.

Practically, the ANN model is simpler in model construction.

The estimation method of both the MVP and ANN models involves stochastic processes. MVP model estimation depends on Monte Carlo Markov Chain method which makes random draws from multivariate normal distribution. ANN model estimation, when using the gradient descent algorithm, relies on the random starting location of each individual weight parameter. To reduce the effect of randomness, ten replications runs are conducted for each model-scale pair. The variance of the computation time and prediction hit rate are evaluated.

The variance of computation time and variance of prediction hit rate among ANN replication runs are higher than that of MVP model runs. For ANN replication runs, variance of computation time is very high as the scale goes up from four to eight and 16. This finding indicates that replication runs are necessary for properly applying ANN model to solve business problems because randomness has a fairly large impact on ANN model's estimation. In contrast, the MVP model has relatively stable computation time and very tiny variance of prediction hit rate. This result indicates that replication runs for the MVP model does not add much value.

The second objective is to apply ANN technique into large-scale cross effect analysis. Instead of simply adopting classic ANN algorithm, this study customizes the ANN technique to make it fit to this specific business problem. For example, the cross entropy function is customized to weigh false negative error more over false positive error because missing a potential customer (false negative) hurt a marketing action more than misidentify a non-customer (false positive). However, existing ANN research in the business field mainly focuses on comparing prediction performance of ANN with that of statistical model. This study goes deeper into ANN model's mechanism and seeks to adapt its learning method to specific business problems.

The last contribution of this study is the introduction of spending habit heterogeneity model. Consumers' heterogeneity in the cross effect literature has not been studied in depth. There are few, if any, dedicated studies. Dedicated heterogeneity studies are very few, if not none. Consumer demographic and historical shopping basket information are used as explanatory variables to

consumers' heterogeneous response in cross effect model. For example, Manchanda et al. (1999) regress effect/cross effect of marketing mix on family size and total number of shopping trips. Russell et al. (1999) estimate the effect of consumers' average basket size on cross effect. Duvvuri et al. (2007) consider fixed effect of income and household size on latent utility. This study focuses on consumer's inherent spending traits. Drawing on the consumer mental budgeting model (Heath and Soll 1996, Duvvuri et al. 2007) and consumer impulsive spending literature (Rook and Fisher 1995, Vohs and Faber 2007), the spending habit heterogeneity model is introduced. This study provides testable propositions based on the model, but leaves empirical testing for future research.

1.2. Theoretical and Managerial Implications

The large-scale cross effect analysis research is constructed in accordance with the business analytics paradigm. Business analytics has been defined "evidence-based problem recognition and solving" (Holsapple et al. 2014). This study, in accordance, incorporates large number of simultaneous categories into model, and (2) leaning on data-driven analysis by relaxing dependence on subjective prior assumptions of possible cross effect.

The big data analytical approach is attracting increasing research interest in the business literature. A recent example is the application of adaptive modeling. This modeling approach, instead of specifying a fix distribution of dependent variables, adjusts distribution specifications based on results of data tests. Cross

effect literature has used machine learning (Mishra et al. 2014) and Dirichlet process (Li and Ansari 2013) as a component of adaptive model calibration. Instead of manually specifying model construct, both Mishra et al. (2014) and Li and Ansari (2013) allow varying the model's hyperparameters – the parameters defining the model construct itself.

These adaptive models are shown as more capable of fitting an information-rich dataset. However, one practical problem with these models is the significantly increased complexity in model specification and estimation. High complexity can impede adoption of the adaptive modeling method and, in turn, slow down productivity gains from utilizing big data analytics. In general, there is an emerging request for large-scale analysis techniques.

In a retailing context, because of the large-scale and high frequency of transactions, a small improvement in accuracy of estimating sales boosts can have significant impact on business operations and bottom lines. For example, a manager of Walmart could find out that a discount of frosting is not necessary when cake mix is on sale because consumers who buy the latter will most of the time also buy the former anyway. Avoiding such a discount can mean a notable profit gain in large volume sales. A significant profitability improvement by a customized discount is demonstrated in empirical studies such as that of Duvvuri et al. (2007). In contrast, when cross effect is ignored, managers can make misleading inferences about the impact of marketing mix (Russell and Petersen 2000, Duvvuri et al. 2007).

Ideally, managers can simply load all of the categories (for example, all categories in a food section) into a large-scale model and the model can identify and quantify cross effects among these categories. By relaxing dependence on prior assumptions, this study is able to explore cross effect between novel combinations of categories. For example, by including all categories in the product group of laundry supplies, managers may find that detergent purchase is influenced by promotion of dryer sheets but not much by promotion of liquid softeners. This is a made-up example but using prior assumptions to speculate cross effect in a set of more than 20 categories is not only difficult but risky. The first premise of large-scale cross effect analysis is to relax prior assumptions about cross effect partners.

1.3. Related Research

This section distinguishes the current study from several related research streams.

1.3.1 Association Rule Mining

In the data mining research field, the well-known association rule mining technique is rooted in finding frequent patterns. Frequent pattern of co-purchase is one example of such patterns. Some literature call it *frequent pattern mining*, or *market basket analysis*. The Support-Confidence framework is the cornerstone of association rule mining research (Agrawal et al. 1993, Agrawal and Srikant 1994, Kotsiantis and Kanellopoulos 2006, Han et al. 2007). Both the measure of *Support* and *Confidence* are a type of frequency measure. This study's research topic, the

cross effect analysis, is different from association rule mining from two perspectives.

First, association rule mining literature focuses on finding purchasing associations based purely on purchase frequency. Its purpose is to identify which two products are highly frequently purchased together. Association rule mining strives to reveal “interesting” purchasing associations. For example, it will drop the association of cake mix and frosting from the resulting rule list because people already know this association and, thus, it is not “interesting”. In contrast, cross effect analysis looks into understanding why and how consumers purchase them. For example, it can specify that consumers use them together and discount on frosting boosts sales of both. Then, it asks the question of how much sales managers can reasonably expect from a certain amount of discount.

This dissertation connects these two research streams. Association rule mining can find interesting rules, but does not explain reasons and does not quantify cross sales. Existing cross effect literature does the opposite, explaining and quantifying but does not look for unexpected associations. This dissertation looks at an integrated capability of rule finding and quantifying. Section 1.2 explains why such an integrated capability is important.

Second, from the methodology perspective, association rule mining focuses mainly on improving computational performance because finishing computation within a bearable time is still a main issue. When working on a large transaction

database, the rule searching is very time consuming. However, the computation time, at least in the published literature, is not a focus of cross effect research.

1.3.2 Affinity Analysis

Affinity analysis carried out in the marketing literature is similar to the association rule mining research in that it aims to analyze purchase associations (Russell et al. 1999, Boztuğ and Reutterer 2008). In general, marketing literature points out that affinity analysis ignores marketing mix and consumer heterogeneity and, thus, may be too misleading to be used in marketing decision making (Russell and Petersen 2000).

1.4. Summary of Research Question

In summary, this dissertation examines solutions and related concerns to large-scale cross effect analysis. Generally, it has two objectives:

1. To examine computation and prediction performance of a MVP model in increasing scale of categories.

2. To apply and customize ANN technique in large-scale cross effect analysis.

Additionally, by synthesizing the existing literature of consumer mental budgeting and impulsive spending, this study conceptualizes the spending habit heterogeneity model in the cross effect context. The model is introduced in the future research section. Testable propositions are provided.

Chapter 2. Background and Literature

2.1 The Random Utility Model

Most cross effect studies take the approach of random utility model (McFadden 1973; 1980; 1986). Walker and Ben-akiva (2002) provide review and generalization of this model. This model specifies that consumers implicitly calculate a utility gain on each transaction. When having an opportunity to purchase a specific category on a shopping trip, consumers do purchase if the utility gain is positive, and do not purchase at a zero or negative gain. This latent utility specification translates consumers' discrete choice decisions into a continuous variable of latent utility. Then this continuous latent utility ($\boldsymbol{u}_i$) can be regressed on interesting independent variables ($\boldsymbol{x_i}$) such as price and promotion. In general, the random utility model can be shown as below.

$$\boldsymbol{y} = \{y_1, y_2, \dots, y_K\}, choice\ of\ K \text{ categories in a COPI set} \tag{2.1}$$

$$y_i = \begin{cases} 0\ (not\ purchased), & u_i \leq 0 \\ 1\ (purchased), & u_i > 0 \end{cases} \tag{2.2}$$

$$u_i(\boldsymbol{x_i}) = \boldsymbol{\beta}_i \boldsymbol{x_i} + \varepsilon_i \tag{2.3}$$

$$i \in \{1,2 \dots K\}$$

Note: For simplicity, the general model shown here has omitted the index of household *h*, and shopping trip *t*.

Consumers' choice of purchase among a COPI set is represented by the vector variable $\boldsymbol{y}$ as shown in equation (2.1). Each element $y_K \in \boldsymbol{y}$ is a binary

valued variable indicating purchased or not purchased of category *K* as shown in equation (2.2). The conditions of equation (2) shows the mapping from discrete choice to latent utility. In short, the model specifies that consumers will purchase category *K* when $\boldsymbol{u}_K > 0$ and vice versa. Equation (2.3) represents a multivariate regression model. Utility of purchasing a category $\boldsymbol{i}$ is regressed to $\boldsymbol{x}_{\boldsymbol{i}}$, a vector of independent variables such as price, product display, product featuring, and other promotions (including the constant term 1 as the first element). The estimated effects are captured in vector $\boldsymbol{\beta}$. The error term captures the unobserved utilities of purchasing category $\boldsymbol{i}$.

2.2 Existing Models of Cross Effect

2.2.1 Utility Correlation and Multivariate Probit Model (MVP)

Manchanda et al. (1999) specify cross effect as the effect of category A's marketing mix on category B's purchase utility. In their MVP model, the latent utility of a customer purchasing a category is regressed on the marketing mix variables of two sources, the focal category and paired category. The latter captures the cross effect, i.e., the part of purchase utility allocated to other categories' attributes. The unobserved purchase utility is captured by the error term. The specific feature of the MVP model is that it allows correlation among error terms. The unexplained co-purchase incidence is captured by the correlation matrix of error terms. The model is estimated with two pairs of categories, cake mix and frosting and detergent and softener. Their model finds that cross-effect driven by marketing mix (price and promotion) is as high as 0.2 for the two pairs of categories.

Li et al. (2005) extend Manchanda et al. (1999) and model choices of financial products. Consumers' readiness to buy a financial product is called maturity status. They theorize that consumers' maturity status can be explained by variables such as cumulative purchases, average account balance, and experience with a type of product. Correspondingly, a financial product can have its maturity level indicating fit to different levels of consumer maturity status. By adding such a set of explainers into the Manchanda et al. (1999) model, Li et al. (2005) show an improved prediction accuracy on a holdout sample. It is also acknowledged that Manchanda et al. (1999)'s model outperforms several alternative models in this context.

Duvvuri et al. (2007) use a model similar to that of Manchanda et al. (1999). They simultaneously load into their model six categories including the four used by Manchanda et al. (1999). But the cross effect partners are pre-specified and are not allowed to change during the model estimation. The results show that some consumers seem more sensitive to spaghetti's price than to sauce's price under an independent model, but they become more sensitive to the price of sauce under a cross effect model. Such an inverse relationship is surprising. The authors explain the results with a mental budgeting model (Heath and Soll 1996).

2.2.2 Conditional Choice and Multivariate Logistic Model

Russell and Petersen (2000) model cross-category incidence with a conditional choice model. The model assumes that consumers' latent utility of choosing a category in a shopping basket depends on the categories that are already chosen. If a cross effect exists, then the data are supposed to show more

frequent purchases of such a combination than purchases without that. Interestingly, their analysis shows that the size of the cross effect (by price change) is pretty small among the four categories they use (paper towels, napkins, facial tissue, and toilet tissue.) Their study suggests that cross price effect exists, but may not be managerially important because of the small effect size.

Their model is called a conditional choice model in the sense that the concept of cross effect is conditioned on an actual purchase of the cross partner (recall appendix A.1 for the model specification). Their model theorizes that the cross effect is the extra utility from purchasing another category, given the current category is already purchased. This dissertation takes the alternative perspective that cross effects do not necessarily rely on actual purchase (recall section 2.4). For example, suppose that consumers purchase both cake mix and frosting when frosting is on sale; but do not purchase cake mix when frosting is not on sale. Russell and Petersen (2000) model captures cross effects from only the former scenario, because the cross effect is the extra utility for purchasing frosting. The model of this dissertation captures cross effects from both scenarios because cross effects could exist without actual purchases. This perspective is less constrained and allows "informational" cross utility. For example, a promotion of flowers may just remind consumers to purchase a box chocolate without purchasing the flowers. The Russell and Petersen (2000) model excludes such cases.

Song and Chintagunta (2006) extend the Russell and Petersen (2000) model by accounting for demand of competing brands within each category. The

result of data analysis, in general, validates the necessity of accounting for cross effect in understanding purchase relationships. At the cost of increased model complexity, the Song and Chintagunta (2006) model demonstrates an approach of directly quantifying the relationship of purchasing among specific brands. The results of data analysis show some unexplained observations. For example, the data model estimation shows that lowering the price of Tide powder detergents increases the sales of liquid detergent.

Boztuğ and Hildebrandt (2003) adopt the model of Russell and Petersen (2000) and test with a German dataset. They find a similar level of cross effect, but an opposite direction of the effect of customers' average basket size on latent utility.

2.2.3 Consumption Satiation Model

Kim et al. (2002) propose a model that is rooted in the consumer theory of micro economics. The model approximates consumers' choice of a bundle of yogurt products with the principle of consumption complementary/substitution, consumer budget constraints, and consumption utility satiation. The model has a stronger theory base and is able to model quantity choice with the utility satiation theory. However, it puts more constraints on choosing complementary/substitute categories. In contrast, this study relaxes the dependence on prior assumptions for what categories being chosen.

2.2.4 Studies Using an Extended Number of Categories

Chib et al. (2002) examine twelve categories that are deemed as composing a classic market basket. Compared with this dissertation, their study does not relax the dependence on prior assumptions for choosing related categories. Moreover,

their study quantifies the correlation of utility of purchase, which is different from our goal (i.e. to quantify the utility dependency) They find that ignoring cross effect (measured by utility correlation) can bias estimation of marketing mix, and using a subset of twelve categories can bias estimation of cross effect. The estimation bias issues is acknowledged in literature, such as (Duvvuri et al. 2007, Hruschka 2013).

Boztuğ and Reutterer (2008) combine two techniques: cluster analysis and cross effect analysis. Cluster analysis theorizes that consumers with similar basket compositions would be similar in latent-utility-based decision making processes. It first clusters shopping baskets by an extended *K-means* algorithm, and outputs 14 basket prototypes, each containing 5 categories. When consumers are grouped into basket prototypes, the whole data set can be divided into subsets, and cross effect models can be estimated on each of these subsets, as well as on the whole dataset. The cross effect model they use is the same as the Russell and Petersen (2000) model.

Artificial intelligence technique is rarely used in cross effect analysis. Hruschka (2014) examines a technique called restricted Boltzmann machine in this context. They are able to load 60 categories into that machine.

In terms of scale of categories, Table 2.1 shows some representative literature and the scale they used.

Table 2.1 Selected representative literature of cross effect

Paper	Model	categories	Data selection method	Research opportunities
(Manchanda et al. 1999)	MVP	{Cake mix, frosting} {Detergent, softener}	2 pairs of categories -> 205 household (2 purchase) -> 155 random -> 17,389 trip made -> 3,414 purchased	* Performance on high dimension COPI is not tested * Household heterogeneity across categories is proposed for future research
(Russell and Petersen 2000)	MVL	{Paper towel , toilet tissue, facial tissue, paper napkins}	4 categories -> 170 households -> 2,578 trips	* Performance on high dimension COPI is not tested
(Kim et al. 2002)	Utility saturation	{Strawberry, blueberry, Pina Colada, Plain, Mixed Berry}	5 categories -> 332 household -> 2,380 trips	* Performance on high dimension COPI is not tested
(Duvvuri et al. 2007)	MVP	{Spaghetti, Sauce, Detergent, Softener, cake mix, Frosting}	6 categories -> 226 household -> 126 random -> 16,032 trip -> 1,656 purchased	* Research is needed to simultaneously model large number of categories
(Hruschka 2013)	Clustering + MVL	Clustering 28 categories on similarity and divided into four groups	Random 1,500 households in IRI database -> 28 categories (7 X 4 groups in analysis) -> 24,074 trips	* Needing more comparison of performance and validity test between small and large number of categories
(Niraj et al., 2008)	MVL	{bacon, egg}	1 pair of categories -> 883 household (5 bacon purchase) -> 467 household (4 egg purchase) -> 293 restricted -> 42,274 trips	* Performance on high dimension COPI is not tested

2.3 Modeling Large-Scale Cross Effect

2.3.1 Parameter Explosion

Relaxing dependency on prior assumptions leads directly to a parameter explosion problem. If category partners are pre-paired for cross effect, such as bacon paired with egg, or cake mix paired with frosting, then parameter explosion can be avoided because a model takes only two categories. If pairs are not pre-specified, then a model needs to quantify cross effects between any possible pair of categories in a set. This results in a set of relationships between categories of any subset. The resulting relationships can be a permutation operation on the original set. By limiting the relationships to pairs only, there are p_{32}^2 = 992 relationships for a set of 32 categories. If there are two independent variables for a category, such as price and promotion, then the number of parameters doubles. This is the case for our extracted dataset. See Section 3.3 for details of dataset preparation. The model's performance with increasing category scale has not been examined in the existing literature.

2.3.2 Reliability of Parameter Estimation

Research finds that co-purchase correlation is identified as insignificant in a small scale model, while weakly significant in a larger scale model (Boztuğ and Reutterer 2008) and the effect size is sometimes underestimated (Chib et al. 2002). Chib et al. (2002) examine the potential impact by including 12 categories, which are identified by prior literature as a classic consumer basket composition. First of all, they found that cross-category effect (they call it cross category correlation) exists and ignoring it will lead to overestimation of marketing effectiveness. More importantly, they found a biased estimation of cross effect on a model of two

categories compared with a model of 12 categories. Their paper is among the few empirical studies that look into the impact of increasing number of categories. The biased estimation found under small number of categories makes it valuable to further investigate impacts of number of categories. More studies are needed for further and more deeply understand the impact of category scale on model performance.

Hruschka (2013) compares the performance of a holistic model of 28 categories with that of four individual models. Each individual model contains 7 categories. The results show that the four individual models generate biased estimations. Additionally, the individual models show insignificant purchase correlations that are shown to be significant in the holistic model. The models focus on the Pearson correlation measure rather than cross effect.

2.3.3 Complexity and Effectiveness of Parameters Estimation

Specifying an econometric model is time consuming and intellectually challenging. It involves carefully designing a data collection/preparation method, plus making reasonable assumptions of data distribution and relationships between variables. Large-scale cross effect analysis apparently increases such complexity. First, it needs to deal with big dataset transformations usually at the level of millions of records. It would be a difficult experience without large-database skills. However, academic business research has a tradition of focusing on theoretical development. The data transformation, cleaning, and selection processes usually draw very little attention.

In terms of parameter estimation efficiency, the multivariate econometric model used in cross effect analysis is subject to the "curse of high-dimensionality" (Manchanda et al. 1999). The MCMC estimation method is commonly used to remedy the difficulty of computing high dimension integrals (Russell and Petersen 2000, Duvvuri et al. 2007, Hruschka 2014). In general, the MCMC method avoids calculating high dimensional integrals by drawing random samples from posterior distribution of parameters. The existing literature shows that the MCMC method works fine in small scale cross effect models. However, performance of MCMC in large dimension problems has not been examined. Ceperley et al. (2012) speculate that the method may be less effective for high dimension problems. This dissertation empirically examines MCMC's performance in high dimension cross effect.

One goal of this dissertation is to examine model performance of MVP for large-scale categories. It also examines performance of artificial neural network model for large-scale categories. Details of the model specifications and study design are provided in Chapter 3.

Computation time may modestly, linearly, or exponentially increase with the increasing number of loaded categories. Information about this relationship has theoretical and empirical meanings. First, if the computation time turns out prohibitive, then it intrigues researchers to find out reasons and to provide solutions. Even if the increase turned out linear or modest, the total computation time may be significant because of the large number of categories loading. For example, it may be affordable to compute cross effect for four categories in four hours, but it

would be prohibitive to compute that for 100 categories in 100 hours, even though the time increase is linear.

Future research may be warranted to improve estimation algorithms and improve computing performance so that managers with low profile computing power can adopt and benefit from large-scale analysis i.e., the commoditization of computing large-scale cross effect. An IBM Whitepaper[1] predicted that, through 2015, 85% of Fortune 500 organizations will be unable to exploit big data analytics for competitive advantage. If analytics capability became an enabling component in large organizations, then middle or small ones have to build that capability to survive. Researchers can contribute to the commoditization of big data analytics. In the area of large-scale cross effect analysis, improving estimation algorithms and computation performance are always significant contributions.

2.4 Cross Effect Concept Development

Manchanda et al. (1999) defines it as the effect of marketing mix of one category on latent utility of another category. This specification has high managerial implications because managers can manipulate marketing mix. This definition, on the other hand, does not explain why price change of category A will affect sales of category B.

This study specifies cross effect as the dependency of perceived utility of purchasing a category on that of purchasing another category. In general, this

[1] IBM Whitepaper (2012) "Business analytics and nexus of information"

specification first provides a broader range of cross effect implementation and, thus, allows more flexible model specification. The utility dependency specification does not contend that marketing mix of A directly impacts sales of B. Promoting A increases A's latent utility. If B's latent utility is conditioned on A's, then B's utility may increase or decrease depending on the strength and direction of the latent utility dependency.

This specification is theoretically justifiable. A consumer's perceived utility of purchasing a category is not only related to the category's own marketing mix, but also related to the consumer's perceived utility of purchasing related categories. For example, many people when making a cake want to use frosting as well. In this case, the utility of purchasing frosting is conditioned on the utility of purchasing cake mix. If consumers do not have positive utility of purchasing cake mix (not plan to make cake), then their utility of purchasing frosting may significantly decrease. Theoretically, such a utility dependency, if it exists, should be reflected in observed dependency of decisions of purchasing.

The root cause to cross effect involves a question of why consumers purchase some categories together under a statistically significant frequency. Russell and Petersen (2000) articulates the two schools of theory, namely, store traffic and global utility. Store traffic theory attributes cross-category incidence mainly to store-specific features such as promotions, displays, and bundles. It assumes that an individual consumer's latent utilities from purchasing each category is independent from each other. Thus, purchase correlation is only observable at the store level, not at the consumer level. The global utility theory

suggests that cross-category purchases are results of a consumer's preference of joint-consumption. So, the cross-category purchase is independent of shopping store, but is supposed to be subject to consumer heterogeneity. Recent literature has found significant consumer heterogeneity in empirical studies. Further, more and more papers take into account consumer heterogeneity in modeling cross-category purchase. These two theories do not necessarily exclude each other in nature, but managers are interested in knowing consumers' heterogeneity. This paper takes the global utility view that takes consumers' consumption utility as the major cause of cross-category purchase.

2.5 Consumer Heterogeneity and Spending habit

Accounting for consumer heterogeneity is necessary in many marketing analysis situations (Wind 1978, Kamakura and Russell 1989, Wedel and Kamakura 1998). Under various conditions, the result of an analysis can be biased or fail to identify expected relationships when heterogeneity is not properly addressed (Allenby and Rossi 1998, Fiebig et al. 2010, Dippold and Hruschka 2013). It is a common practice to incorporate consumer heterogeneity in cross effect research (Duvvuri et al. 2007, Niraj et al. 2008, Mehta and Ma 2012, Aguinis et al. 2013). However, heterogeneity in cross effect studies typically is correlated simply with demographic information or historical transaction information. Theoretical investigations in the cross effect research is very limited. One exception is that of Duvvuri et al. (2007) which examines consumer budgeting theory for possible causes of heterogeneous response in cross effect.

Consumer heterogeneity is related to subjects such as market segmentation, conjoint analysis, and cluster analysis for the similar objective of distinguishing consumer preferences. This study first reviews the heterogeneity studies in cross effect literature. Then, it provides a quick overview of market segmentation literature for segmentation methods, variable selection, and evaluation criteria. Finally, this investigation articulates the heterogeneity model used in this dissertation.

2.5.1 Heterogeneity in Cross Effect Literature

There are mainly two types of heterogeneity models in cross effect literature. One type is represented by Chib et al. (2002) and the other is exemplified by Manchanda et al. (1999).

Chib et al. (2002) capture household-specific heterogeneity and category-specific heterogeneity with a fix effect model. Their model assumes that each household has its own mean utility on each category. The equation (2.4) below shows the general model. See Chib et al. (2002)'s model in the Appendix A.1.

$$U_{htj} = X_{htj}\beta_j + b_h + c_{hj} + \varepsilon_{htj} \tag{2.4}$$

The utility, U_{htj}, is attributed to household value b_h, and household-category specific value c_{hj}.

The model implicitly assumes that households are the same in elasticity of marketing mix. The parameter β_j has a one-dimension index only on category *j*, not on household *h*. It indicates that the effect of marketing mix of category *j* is

same across all of households. Such a fixed elasticity specification is relatively arbitrary because it suggests that a change of marketing mix of a category generates a utility change that is the same for all consumers. This specification rules out the possibility that consumers perceive different values of a same change of marketing mix.

Manchanda et al. (1999) fit a random effect model of heterogeneity. The model assumes that household specific cross effect is normally distributed. For a given household, the cross effect value on a category is determined by individual characteristics, plus a random error. In segmentation literature, this approach is sometime called mixture model (Kamakura and Russell 1989, Wedel and Kamakura 1998). The consumers in the data sample are assumed to be a mix of different groups. Each group of consumers is assumed as a random draw from the "super" population. Thus, a group has its own distribution of the cross effect, which may be featured by different mean and variance derived from the "super" population's overall mean and variance.

Both Duvvuri et al. (2007) and Manchanda et al. (1999) take purchase history as an independent variable, but in different ways. The former includes the inventory variable at the level parallel to marketing mix (i.e., the explanatory variable of latent utility). The latter considers the purchase history (purchase frequency) at the level parallel to demographic variables (i.e., the explanatory variables of the heterogeneity). The latter approach is consistent with market segmentation literature (Allenby and Rossi 1998). Information of consumers' demographics, social status, and behaviors are better understood when integrated

into abstraction of lifestyles (Holt 1997) instead of being used directly to explain purchase utility.

Appendix A.1 summarizes the literature in regards to ways of dealing with heterogeneity in cross effect models. It should be noted that the network models have been examined for addressing heterogeneity (Yang and Allenby 2003).

Consumer heterogeneity is highly related to the market segmentation research stream, which focuses on addressing consumers' different preference. Depending on the research objectives, there are many ways to segment consumers. This study briefly reviews the segmentation approaches in the next section.

2.5.2 Segmentation Method in Marketing Literature

In the marketing literature, segmentation research is about theories and methods to capture heterogeneous consumer needs and preferences. Such efforts in general enable marketers to better identify and serve customers with precise customization. The value of segmentation has been well acknowledged in both academic and industrial marketing research and practices. For example, Currim (1981) find that by segmenting customers by their perceived utilities on transportation alternatives, analysts can identify important factors that are specific to a segment in terms of influencing their choice of transportation alternative. These factors would not be identified as important in an aggregated model.

Market segmentation is rooted in heterogeneity of consumer needs/preferences. Smith's (1956) definition was cited by Wedel and Kamakura

(1998, page 1), which says “market segmentation involves viewing a heterogeneous market as a number of homogeneous markets, in response to differing preference, attributable to the desires of consumers for more precise satisfaction of their varying wants” and also says “…segments are directly derived from heterogeneity of consumer wants…”

Market segmentation is a relatively mature subject in marketing literature (Wedel and Kamakura 1998, page 1, Taylor-West et al. 2008). Classic articles include, but are not limited to, Smith (1956), Wind (1978), Punj and Stewart (1983), Kamakura and Russell (1989), Jedidi et al. (1997), Allenby and Rossi (1998), Straughan and Roberts (1999), Boxall and Adamowicz (2002). Through decades of literature accumulation, market segmentation has developed a rich set of methods and models. Thus, this dissertation relies on this literature in developing its heterogeneity construct.

This study excludes literature on conjoint analysis (Green and Srinivasan 1978) and related studies that are based on product attribute utility models. Conjoint analysis is recognized as a methodology for developing market segments (Wedel and Kamakura 1998, chapter 17). It theorizes that a consumer’s utility of purchasing a product is based on a function (usually linear) of the consumer’s evaluation of the product’s attributes. In contrast, the cross effect is theorized on the utility dependency between categories and, thus, is a higher level aggregation than conjoint analysis. In terms of addressing consumers’ heterogeneity in cross effect, it is often theorized that consumers are essentially heterogeneous, but the

heterogeneity can be aggregated according to research objectives and data availability.

Punj and Stewart (1983) provide theoretical justification for applying cluster analysis to solving marketing problems. In general, segmentation and cluster analysis share a goal of grouping entities such as consumers and companies. Practically, cluster analysis is traditionally recognized as a method of market segmentation (Punj and Stewart 1983, Wedel and Kamakura 1998, page 17).

2.5.3 Spending Habits, the Heterogeneity in This Dissertation

This study conceptualizes spending habits as a source of heterogeneous response to marketing mix. This conceptualization is based on two schools of literature: mental budgeting model (Heath and Soll 1996, Duvvuri et al. 2007) and impulsive spending research (Rook and Fisher 1995, Baumeister 2002, Vohs and Faber 2007). The mental budgeting model suggests that people conduct implicit calculation of purchase utility and make decisions when the calculation results in obvious gains/loss of a purchase.

This study takes the mental budget model further and considers the depth of consumers' mental calculation. People of careful spending habit may be accustomed to planned and within-budget spending. Thus, they are more likely to do deeper calculation on purchases. At one end of the spectrum are people who rarely spend over/under budget, while the other end has people who care little about a price tag, but buy things catching their attention. Impulsive spending research has a viewpoint backing this idea. The viewpoint suggests that people

can be inherent “impulsive buyers” or “not impulsive buyers “(Rook and Fisher 1995, Youn and Faber 2000). This kind of spending habit (depth of mental calculation) is expected to be stable because habits are very hard to change.

Cross purchasing is a phenomenon that is expected to be highly related to consumers’ budget calculations. In general, at a promotion or price drop, consumers buy more. In a case where cross category is considered, consumers may have to buy more of both to enjoy the utility of discount or promotion. Whether the cross category would be purchased can be associated with flexibility of her mental budget, i.e. the spending habit. Indulgent types of mental budgeting are hypothesized to be associated with higher cross effect, because the consumer’s mental budgeting is allowed larger variance.

To realize the mental budgeting theory in cross effect model, the first task is to define variables that reflects consumers’ types of spending habits. Plausible variables should reflect consumers’ degree of mental budgeting. It is assumed that the total spending at shopping trips is normally distributed. Thus, large variance of trip total spending can reflect indulgent mental budgeting, while small variance can reflect conservative mental budgeting. Technically, the variance of trip total spending is easy to calculate. The value range is (0, +Inf).

It is possible that the normality assumption is violated. For example, consumers may periodically make a large grocery shopping trip followed by small contingent shopping. This type of consumer can have large variance of total spending but still be budget sensitive. This study does not expect strong presence

of such cases in its dataset, because the contingent trips would be excluded if the focal categories was not purchased in that trip. Even if the pattern appears in individual consumer's purchasing history, it would not seriously disable our model. The random effect heterogeneity model (recall next section) pools consumers by their degree of mental budgeting, rather than distinguishing individuals. Pooling consumers together largely removes such patterns.

2.5.4 Data Sparseness, Fixed Effect and Random Effect Model of Heterogeneity

This study specifies a random effect model for heterogeneity. The choice between random and fixed effect model is highly related to data sparseness. The two models are explained below.

The fixed effect approach assumes that each unit is significantly different from others in terms of responding to a stimuli, and estimating the unique response of each unit is necessary according to research objectives. In contrast, the random effect approach assumes that, even though each unit is different from another for response to a stimuli, the difference among units is not serious enough to be distinguished between every pair of individuals according to research objectives. Thus, it is good enough to estimate a general center for all units and a unit's "extra" response is regressed on explanatory variables. In this sense, regression models in general are random effect models where individual's response is not uniquely identifiable by the model.

For example, managers can regress sales change on price change. The estimated model can predict sales change caused by a unit change of price. This

model represents a random effect of price on sales because it does not distinguish individuals for their response, but rather estimates a general center over all respondents. In another case, a manager of target marketing wants to estimate consumers' response to a marketing campaign for each individual customer because customized promotion can be developed. In this case, individual personal information must be used to estimate the different responses. In other words, the value of the estimated model lies at the individual's unique feature that makes him/her respond differently; while in the former case, the value lies at the overall aggregate level of response.

The choice between fixed and random effect heterogeneity models depends not only on research objectives and theoretical justification, but is also constrained by data availability. Consumer's response to a marketing mix change can be very sparse. As explained in (Rossi et al. 2005, page 130), 12 observations is very common for a household's purchase of a category in a year. In such a situation, it is necessary to consider the proper level of aggregation, even though the individual level of heterogeneity is desirable.

Data aggregation level is a critical factor for addressing heterogeneity. Theoretically, each consumer is unique in consumption preference. But, each pair of consumers can also share a certain level of aggregated preference. Marketing segmentation literature sometimes labels it segment homogeneity (Wind 1978, Wedel and Kamakura 1998). For example, the well-known Maslow's hierarchy of needs depicts the most abstract level of human beings' needs. People need water,

as well as a feeling of love and belonging, but everybody is different in the way and strength of those needs.

Ideally, marketing science would like to model preference as specifically as possible. The extreme end of this direction is the ability to predict each individual's personal preference (Wedel and Kamakura 1998, page 1). It appears that marketing science is heading in that direction through technology advancement.

In a statistical sense, to model cross effect to a specific level of consumer heterogeneity, the available data are required to contain enough samples for that level of preference. For example, in order to confidently cluster consumers' preferences by the factor of family size, enough samples are required for each type of family size to represent a set of distinguishable preferences. Data availability was pointed out as a potential issue for segmentation validity check (Wind 1978). Theoretically a segmentation study should ensure customer identifiability (Wedel and Kamakura 1998, page 4). A large database requires choosing a proper level of preference aggregation so that important factors are included and preferences are properly addressed. In addition, high specificity of preference requires powerful computation resources. Thus, the practical way is to model the preference at the level that is computationally feasible and includes managerially important factors.

2.6 Artificial Neural Network (ANN) Technique

This section first introduces ANN model and explains why ANN is appropriate in large-scale cross effect analysis. It then reviews several representative studies that are related to cross effect analysis.

2.6.1 Advantages of ANN for Cross Effect Analysis

The ANN model has been compared with regression models, especially with logistic regression models, from a statistical perspective (Bishop 1995, Warner and Misra 1996, West et al. 1997). Many empirical studies suggest superior prediction accuracy of ANN over traditional regression models. Wong et al. (1997) reviews ANN's application in business. Paliwal and Kumar (2009) provides a review of ANN from a technical perspective.

This study examines the ANN model in large-scale cross effect analysis for several reasons.

First, the ANN model is a learning model of high adaptive capability (Cybenko 1989, Hornik et al. 1989). It is well documented that ANN captures both linear and non-linear relationships (West et al. 1997, Paliwal and Kumar 2009). Actually by restricting the number of hidden perceptrons to zero and the activation function to direct linear, a neural network downgrades to a linear regression model; under the same condition, but as a logistic activation function, it becomes a logistic model. Both multiple regression and logistic regression models are special cases of generalized linear models. Further, a generalized linear model is a special case of a neural network model (Bishop 1995). The advantage of ANN is that it can learn model structure from data, in contrast to the traditional approach that analysts are responsible for pre-specifying a model structure. This fact suggests that ANN needs a large and information-rich dataset to be accurate. The successful application of ANN in image recognition and natural language processing is based on rich training samples.

ANN is a learning model where the model's structure is not pre-specified by the user but is learned from data. This feature fits the assumption that large-scale cross effect analysis does not pre-specify which is related to which in a set of categories. Because of loading a large number of categories, the large-scale cross effect analysis is by nature hard to pre-specify between-category relationships. The cross effect is less and less manageable manually when the number of categories increase. Utilizing an ANN model's learning capability is ideal for treating this concern.

Additionally, ANN is able to learn non-linear relationships between IVs and DVs, which is usually regulated in regression based econometric models. Finding the non-linear cross effect provides new information for decision support.

Another reason is related to the parameter explosion issue (recall section 2.3.1). ANN avoids parameter explosion in the cross effect analysis situation. The ANN modeling does not require specifying an explicit cross effects variable. But rather, the cross effect is captured in the hidden layer that connects inputs to outputs. An ideal ANN model requires only inputs and outputs, and the modeling procedures is done by a learning process which is transparent to users. Avoiding parameter explosion is actually a by-product of delegating model specification to an artificial learning algorithm. With a set of learned rules, it can mimic decision makers and make decisions based on novel inputs.

In essence, the training process of ANN resembles training a human being with super brain power who can quickly go over millions of past events and build

decision rules along the way. Just like human beings, a learned ANN model may be able to make accurate predictions without being able to articulate the logic or reasoning behind the decision making. With decades of research efforts, researchers have made progress in the ANN model's interpretability such as the generalized weights technique (Intrator and Intrator 2001), marginal cross effect interpretation of a network (Hruschka 2014), and effect quantification with ANN modeling (Xu et al. 2013)

Finally, model specification and implementation is operationally much simpler on an ANN model than on an econometric model (Paliwal and Kumar 2009). In general, specifying an ANN model is more intuitive and requires less in the way of intellectual investments.

2.6.2 Selected Studies of ANN Application in Marketing Problems

ANN application in marketing research is growing. Several studies focus on applying new techniques in specific contexts. For example, Kim et al. (2005) apply an ANN technique in the customer targeting context. Cui et al. (2006) combine an evolutionary algorithm (an important ANN training algorithm) and Bayesian network to study direct marketing response models. The findings are usually in the form of comparing prediction accuracy of ANN to a base model. This study follows the literature convention, and applies the ANN technique in the large-scale cross effect context which is featured by the multivariate model. Additionally, this study examines effects of customizing ANN model configurations to fit the large-scale cross effect context.

The literature of applying ANN in cross effect analysis is rare, if available at all. To give a basic idea of ANN application in marketing literature, several studies that are close relatives to cross effect analysis are reviewed.

Hruschka (1993) compares the regression method with the ANN technique in the context of market response (i.e., predicting sales by price, advertising, lagged advertising and temperature). The results show that ANN with only one layers of hidden nodes generates error (MSE) much smaller than that of the regression model. It does not focus on cross effect but it showcases comparison work between ANN and regression model.

West et al. (1997) compare ANN with discriminant analysis and logit model in the context of consumer choice model. Both simulated and survey data are analyzed. The results show that the ANN outperforms the discriminant/logit technique when data are non-learn nature; but does not outperform when the data are linear in nature.

Cooper (1999) compares ANN with multivariate statistical models, contending that ANN works better when the relation between DVs and IVs is unknown. This finding supports this dissertation's argument that cross effect analysis is a complex problem and ANN is a good fit to this problem.

More recent literature examines feature selection capability of an ANN model. Kim et al. (2005) examine application of ANN in selecting proper IVs in the context of market response. Xu et al. (2013) examine ANN's feature selection application and test an ANN model in predicting emergency room arrivals.

ANN application in cross effect analysis is not available. This study fills this gap and demonstrates ANN's application in the large-scale cross effect context.

2.6.3 Prediction Model

Note that the ANN model is specially designed for prediction purposes. As compared with regression models, it relaxes the assumptions of linear relationships between DVs and IVs. The logit model actually moves from linear regression to the log ratio transformed linear regression. In the same sense, ANN moves from log ratio transformed linear regression to the multi-layer compounded log ratio transformed linear regression with customizable transformation function. Such a model structure escalation can tremendously improve model capability because it does not assume linear relationships, nor assume specific distribution of DVs and IVs. Costs are (1) loss of relationship traceability, and (2) increase of model estimation complexity.

There are continuous breakthroughs to the latter issue because of the fast development of computation capacity. With small searching effort for good parameters, the classic Newton-Raphson's method (gradient decent) works pretty well in practice. One of the recent breakthroughs is the application of GPU chipset in computing deep neural network that can significantly improve the calculation speed (Bergstra et al. 2011). Distributed deep neural network structure is also attracting research efforts (Dean et al. 2012).

Chapter 3. Model Specification and Study Design

This chapter specifies two models of large-scale cross effect analysis.

3.1 MVP Model with Heterogeneous Spending habit

This study uses a multivariate probit model for large-scale cross effect analysis. As with the approach of Manchanda et al. (1999), the model allows correlation between utility errors. The main feature of the model is that it allows correlation between any pair of categories, rather than just between pre-selected pairs.

This study specifies a two-level hierarchical model. The first level model specifies effects and cross effects of marketing mix on the latent utilities of the k categories. The second level model specifies heterogeneity of the effects/cross effects captured by a consumer's spending habit. Tables 3.1 and 3.2 depicts some naming conventions used in model specification.

Table 3.1 Variable naming conventions used in model specification

Variable form in equation	Explanation	Example
Lower case letter	scalar variable	$\boldsymbol{y}, \boldsymbol{\beta}$
Lower case bold letter	vector variable	$\boldsymbol{y}, \boldsymbol{\beta}$
Upper case letter	matrix variable	$\boldsymbol{\Sigma}, \boldsymbol{V}$

Table 3.2 Variable abbreviation conventions used in model specification

Conventional variable abbreviation	Explanation
IV	independent variable
DV	dependent variable
P	price
M	promotion
U	utility
SPD	spending variance
MVN	Multivariate normal distribution

3.1.1 The First Level of the MVP Model

In a shopping trip, consumers choose to purchase or not purchase on each category in a set of k categories. Let vector $\boldsymbol{y}_{th} = \{y_1, y_2, \dots, y_k\}_{th}$ represents choices on each category made by household h at shopping trip t. The element y_{k} is a binary variable where 1 (or 0) indicates the category k was purchased (or not purchased). For example, given that the interest is in purchases of bacon and eggs, if household h purchased bacon but not egg at trip t, then $\boldsymbol{y}_{th} = \{y_{bacon} = 1, y_{egg} = 0\}_{th}$.

The value of y_{kth} is modeled with a latent utility variable u_{kth} as specified in equation (3.1).

$$y_{kth} = \begin{cases} 0\ (not\ purchase), & u_{kth} \leq 0 \\ 1\ (purchase), & u_{kth} > 0 \end{cases} \tag{3.1}$$

The first level of the MVP model is specified as below (Manchanda et al. 1999).

$$\boldsymbol{u}_{th} = X_{th}\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_{th}\ ; \qquad \boldsymbol{\varepsilon}_{th} \sim MVN(\mathbf{0}, \Sigma) \tag{3.2}$$

Dependent variable (DV) $\boldsymbol{u}_{th}$ is a k-dimesion vector represents latent utilities of the k categories for which cross effects are modeled, so that $\boldsymbol{u}_{th} = \{u_{1th}, u_{2th}, \dots u_{kth}\}$. Because the first level model does not account for heterogeneity, the estimated effect (vector $\boldsymbol{\beta}$), is constant across trips and

household. For simplicity, index h and t are temporarily omitted and the general model has:

$$\boldsymbol{u} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \; ; \quad \boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \Sigma) \tag{3.2+}$$

- Identification problem

$\boldsymbol{u}$ is not observed, but rather it is inferred from the observed variable $\boldsymbol{y}$. Thus, $\boldsymbol{u}$ is essentially treated as a model parameter. The model allows all positive value of $\boldsymbol{u}$ when corresponding category is purchased $(y = 1)$, and allows all negative values when not purchased $(y = 0)$. For a given observation(X, y), the parameter of $(\boldsymbol{\beta}_h, \Sigma)$ cannot be uniquely identified because $\boldsymbol{u}$ is free to change in the range (0, +Inf) or (-Inf, 0). An approach to make the model parameter identifiable is to restrict the error covariance matrix, Σ, to be a correlation matrix (Manchanda et al. 1999, Rossi et al. 2005, Duvvuri et al. 2007).

- Parameter organization

The equation (3.2+) in the dimension of k categories is broken down as shown in (3.3). Note that the breakdown is in the dimension of categories, not the dimension of IVs. In other words, each element $\boldsymbol{x}_i$ in the equation of (3.3) is still a vector (i.e., the holder of all IVs of category i). The element $\boldsymbol{\beta}_i$ is a vector, the holder of coefficients to vector $\boldsymbol{x}_i$.

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & 0 & 0 & 0 \\ 0 & \boldsymbol{x}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \boldsymbol{x}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix} \tag{3.3}$$

The matrix X, the first matrix on the right hand side, is a diagonal matrix. Each element $\boldsymbol{x}_i$ is identical and contains all IVs for the set of K categories. The result is that coefficients vector $\boldsymbol{\beta}_i$ is estimated with same $\boldsymbol{x}_i$ and different u_i. Writing in the form of (3.3) enables each same-value IV vector $\boldsymbol{x}_i$ to pick up a different coefficient vector $\boldsymbol{\beta}_i$. By further extending the term $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_i$, we have the following equation.

$$\boldsymbol{x}_i = [1 \quad (p_1 \quad m_1) \quad (p_2 \quad m_2) \quad \cdots \quad (p_k \quad m_k)] \tag{3.4}$$

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1p} \\ \beta_{i1m} \\ \beta_{i2p} \\ \beta_{i2m} \\ \vdots \\ \beta_{ikp} \\ \beta_{ikm} \end{bmatrix} \tag{3.5}$$

Equations (3.4) and (3.5) show that $\boldsymbol{x}_i$ is a row vector of $2 * k + 1$ dimension, and $\boldsymbol{\beta}_i$ is a column vector of $2 * k + 1$ dimension. The whole $\boldsymbol{\beta}$ vector is the combination of $\boldsymbol{\beta}_i$ for $i = 1 \sim k$. Thus, the whole $\boldsymbol{\beta}$ has dimension $k * (2 * k + 1)$. To model four categories, the number of β is 36. To model 32 categories, number of β is 2,080. Allenby et al. (2005) state that a model with such a large number of parameters was not imaginable short time ago, but computer and Monte Carlo Markov Chain (MCMC) simulation method makes such models a commonplace today.

The deterministic terms can be extended as:

$$\boldsymbol{x}_i\boldsymbol{\beta}_i = \beta_{i0} + \sum_{j=1\sim k}(\beta_{ijp}p_j + \beta_{ijm}m_j) \tag{3.6}$$

When $j = i$, the coefficient of β_{iip} and β_{iim} captures effect of price and promotion on focal category i. When $j \neq i$, the coefficients captures cross effects from category j to i.

Most software packages take advantage of vectorization for computation efficiency in parameter estimation. To feed data to such packages, analysts usually need to prepare data in the matrix form. More specifically, they need to prepare a data matrix of Y and X for a model $Y = X\beta$. The matrix form in (3.3) is derived from the fact that effects of same IVs on different DVs are estimated, and the DVs are assumed correlated and follow a MVN distributon.

3.1.2 The second level of the MVP model

The second level of the MVP model specifies household heterogeneity of the first level's parameters $\boldsymbol{\beta_h}$ using a random effect model.

$$\boldsymbol{\beta}_h = \boldsymbol{hh}_h\boldsymbol{\gamma} + \boldsymbol{\xi}_h \; ; \; \boldsymbol{\xi}_h \sim MVN(\mathbf{0}, V_\beta) \tag{3.7}$$

In such a specification, the first-level model must use household as unit. The second- level model assumes that the vector $\boldsymbol{\beta_h}$ is a multivariate normal distribution and the center is determined by household characteristics. Note that in this model, the unit is household. The DV is $\boldsymbol{\beta_h}$, and IV is characteristics of

household. Index h is omitted and the category dimension is broken down as shown in (3.8).

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{hh}_1 & 0 & 0 & 0 \\ 0 & \boldsymbol{hh}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \boldsymbol{hh}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_k \end{bmatrix} ; \quad \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_k \end{bmatrix} \sim MVN(\mathbf{0}, V_\beta) \qquad (3.8)$$

As with the specification in (3.3), the $\boldsymbol{hh}_i$ vector repeats itself in the diagonal matrix in (3.8).

The interest of the second level model is the parameter $\boldsymbol{\gamma}$ and V_β. The $\boldsymbol{\gamma}$ captures the effect of household characters on the parameter $\boldsymbol{\beta}$. The variance/covariance matrix V_β represents uncaptured variance/covariance of $\boldsymbol{\beta}$.

Cross effect models specification is complex. Representation in a matrix view, as in (3.9), can help us clearly understand the model's structure. More importantly, it helps readers understand data organization, which is critical for estimation procedures.

In the hierarchical model, the second layer IVs, compared with first layer ones, are assumed to be an order of magnitude further from DVs. In other words, it assumes that the second layer IVs influence DVs only through the first layer IVs. The legitimacy of such a specification hinges on scientists' knowledge and/or reasoning processes.

3.1.3 Data Sparseness and Alternative Solution

Data sparseness was discussed in section 2.5.4. Serious data sparseness can invalidate parameter estimation. For example, suppose a household makes 12 detergent purchases in a year. But purchase Information is available for only seven of these purchases. The effect of price drops or promotions is not strongly present in the several purchase transactions. Thus, the estimated effect can be unpredictable. In this situation, the two-level MVP model that depends on individual consumer's shopping history becomes invalid.

This study examines the problem of individual level sparseness by validating the first- level price effect. It is well-acknowledged that, at an aggregate level, price is negatively related to demand. This knowledge has been used to check models' face validity (Manchanda et al. 1999, Russell and Petersen 2000). It first ignores heterogeneity and runs the first-level model only. A negative effect of price on utility is expected. After breakdown is taken into account, the two-level model can be run. Data sparseness is deemed present when the two-level model does not give a stable negative effect of price on utility.

When the data sparseness problem is present, the two-level model cannot be relied on to address heterogeneity. The alternative solution is to aggregate individual level data to a higher level.

3.1.4 Bayesian Inference and Monte Carlo Markov Chain (MCMC) Methods

The MCMC method is thoroughly explained in Rossi et al.'s book (Rossi et al. 2005). A brief review of the key parts of this estimation method is provided below.

In the context of large-scale cross effect analysis, the main advantage of MCMC is that it avoids the difficulty of calculating high dimension integrals. Because latent utility is unobserved in the cross effect model, it is necessary to calculate integrals on possibilities of positive utility for a purchase and of negative utility for a non-purchase. For a detailed discussion of this problem, see (Manchanda et al. 1999).

The MCMC approach first specifies a statistical distribution for each model parameter, called prior distribution or prior. For example, a normal distribution of price coefficient as $N(0.02, 0.01)$ can be used. The mean and standard deviation are usually initialized with random numbers. It turns out that the initial value does not matter when a data sample is large and strong. Observed data are treated as evidence that is used to adjust the prior and derive the posterior distribution or post. When the post is available, random samples can be drawn to estimate parameters' values. MCMC is rooted in Bayesian statistics (Rossi et al. 2005).

Bayesian is a theoretical jump away from ML. If ML seeks maximizing the likelihood function Prob(data | θ), then the Bayesian approach adds a wrapping layer to the likelihood function as:

$$\text{post}(\theta \mid \text{data}) \propto \text{Prob}(\text{data} \mid \theta) * \text{prior}(\theta)$$

ML searches for a "best" θ that maximizes the probability function; while MCMC wrapping layer forms a probability distribution of parameters from which the parameters can be drawn. Drawing random samples from a known distribution is much easier than calculating high dimension integrals.

3.2 ANN Model

This section provides a brief overview of a general ANN model specification. It is followed by a section specifying a cross effect ANN model. Finally, several configuration options are discussed as approaches of customizing ANN to fit the cross effect analysis context and fit specific characteristics of the dataset.

3.2.1 The General Construct of ANN

ANN techniques evolved quickly over decades of development. But the fundamental theory and principles have remained solid. There is classical literature introducing the general ANN model such as Bishop (1995) and Ripley (1996). Literature dedicated to comparison of statistical method and ANN is also readily accessible (Warner and Misra 1996, Dreiseitl and Ohno-Machado 2002, Kumar 2005, Paliwal and Kumar 2009). Application of artificial intelligence in business has been growing (Hruschka 1993, West et al. 1997, Wong et al. 1997, Baesens et al. 2002, Cui and Curry 2005, Xu et al. 2013). Online resources are very rich and updated more frequently to the latest technique. For example, the online book by Michael Nielsen (Nielsen 2015) explains principles and techniques of ANN with vivid examples. The UFLDL Tutorial, contributed by well-recognized ANN

researcher Andrew Ng and his colleagues, covers general ANN models and techniques (Ng 2010).

All the resources mentioned above are very consistent in explaining ANN principles and general ANN models. A general introduction of ANN construct is provided in this section. It adopts two figures from (West et al. 1997) to explain the basic idea of NN models.

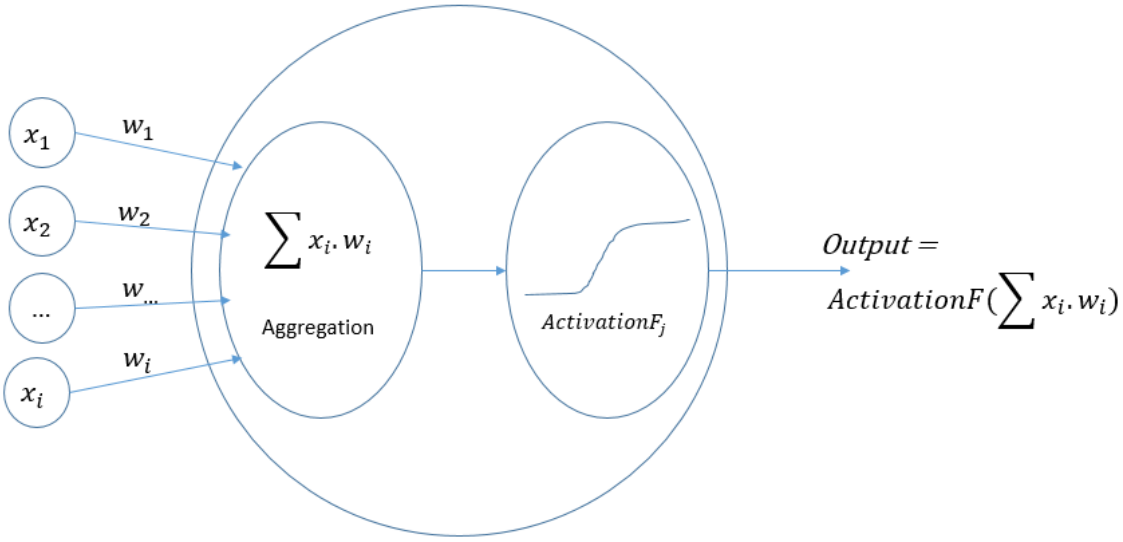


Figure 3.1 Single neuron perceptron, adopted from (West et al. 1997)

Figure 3.1 shows the construct of a single neuron. First, the aggregation node inside the big circle takes I inputs (x_1 to x_i) to calculate an aggregation value. The activation function node takes the aggregated value and transforms it to an output value. In most ANN implementations, the aggregation is a linear combination or weighted summation, which can be noted as $\sum_i x_i w_i$. Parameter w_i is the parameter be optimized using learning algorithm. A commonly used function activation function is sigmoid. Sigmoid function is what econometricians called logistic function. The big circle containing the aggregation and activation function

represents a neuron. Many such neurons can be connected to form a network as shown in Figure 3.2.

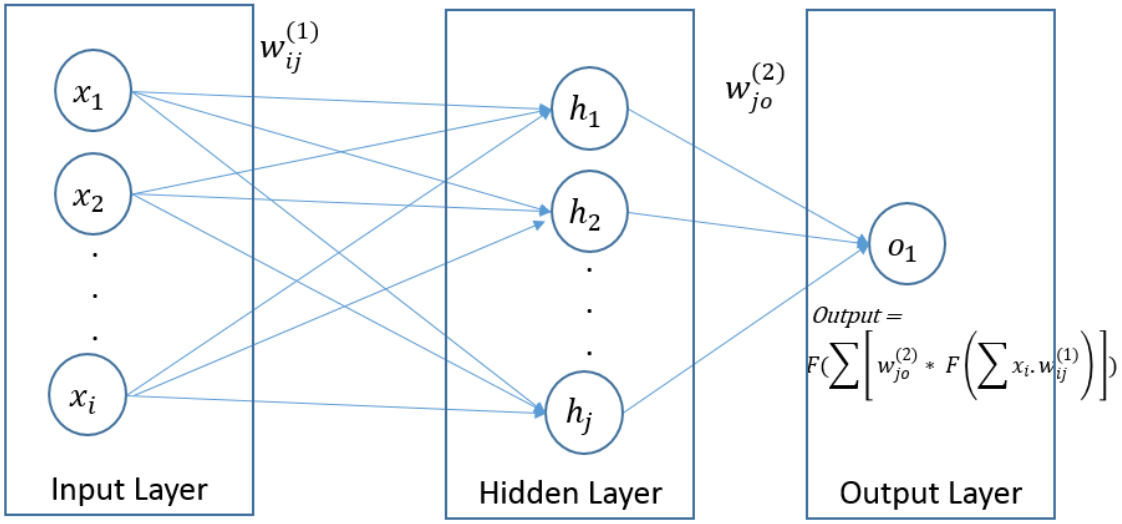


Figure 3.2 One Hidden Layer NN, adopted from (West et al. 1997)

Figure 3.2 shows a three hidden neurons (node H) ANN model. The node O on the right hand side is the output node. The output node can be in the form of a neuron or any aggregation and transformation function defined by the model designer.

Multiple layer NN is not considered here for two reasons. First, one hidden layer is capable of capturing non-linear relationships (Intrator and Intrator 2001, West et al. 1997, Paliwal and Kumar 2009). Second, multi-layer NN is much more complicated in terms of choosing training algorithm and optimal structure design (Bengio et al. 2009, Nielsen 2015 Chapter 5). For example, it may be more subject to the trap of local optima (Baczyński and Parol 2004), especially when gradient descent learning is used. This project focuses on the one-hidden layer ANN model

for cross category predictions. A question remaining is how many hidden nodes should be chosen. This question is discussed in Section 3.2.4.

The figure below is the conceptual decision making model in cross category context.

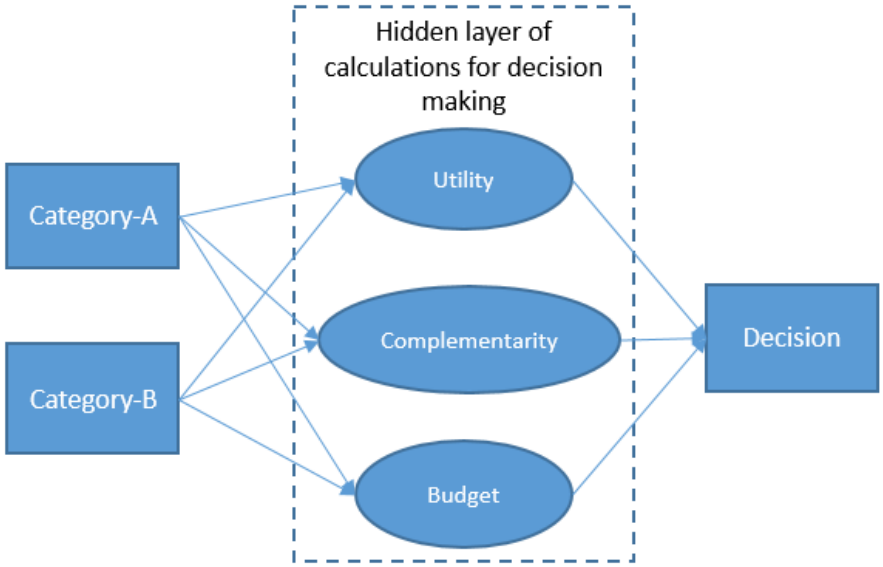


Figure 3.3 Conceptual decision making model

To train an ANN model, one needs to specify a learning algorithm, error function, and stopping rule. The learning algorithm is a process to search for a set of optimal parameters that minimize the error function. There is no deterministic solution to ANN model. Thus, the learning is in essence a repeating "search" and "test" process. The main parameters are the weights, W_{ij}, as shown in Figure 3.2.

- Model Initialization

Common practice is to initialize the weights as random numbers drawn from range 0 to 1.

- Learning algorithm

A learning algorithm takes many rounds to update parameter (weights) toward an optimal solution. Back-propagation is arguably the most commonly used algorithm. In each round, it feeds data to the ANN model and calculates an error with current parameters. Then the parameters are updated using a searching rule. In the next round, the new parameters are used to calculate the error. Gradient descent (Newton's method) is usually used to update parameters in each round. Equation (3.9) describes the basic gradient descent method.

$$w_{ij}^{t+1} = w_{ij}^{t} - \eta \frac{\partial E}{\partial w_{ij}^{t}} \tag{3.9}$$

In round t, a set of weights w_{ij}^{t} is either learned from previous round or set as the initial value when $t = 1$. E is the error function. $\frac{\partial E}{\partial w_{ij}^{t}}$ is the first derivative of E over a weight parameter w_{ij}^{t}. This term represents the gradient descent, the change of w_{ij}^{t} leading to largest reduction of error function. The new w_{ij}^{t+1} is calculated using formula (3.9). The learning rate parameter η controls for the speed of parameter updating. A convenient feature of the gradient descent learning formula is that the derivative term $\frac{\partial E}{\partial w_{ij}^{t}}$ is usually a known form depending on the error function. Thus, it is calculated in each round and used in next round.

- Error function

The learning algorithm is designed to minimize an error function. An error function is a quantity to measure the difference between true value and estimated

value of dependent variables. A commonly used error function is sum of squared error and cross entropy error.

- Stopping rule

The learning algorithm stops when either the solution converges to a point where further rounds of learning would not gain much improvement on reducing error; or, the predefined maximum number of rounds is reached without solution convergence. The former case results in a success while the latter is a failure of learning.

3.2.2 Cross Effect ANN model

Figure 3.4 shows the cross effect ANN model. The input layer includes nodes of dependent variables (category price and promotion). A 32 categories cross-category model includes 64 input nodes. The hidden layer is a one-layer set of neurons including a bias node. The choice of the number of hidden nodes is discussed in a later section. The output layer has a number of neurons corresponding to number of categories.

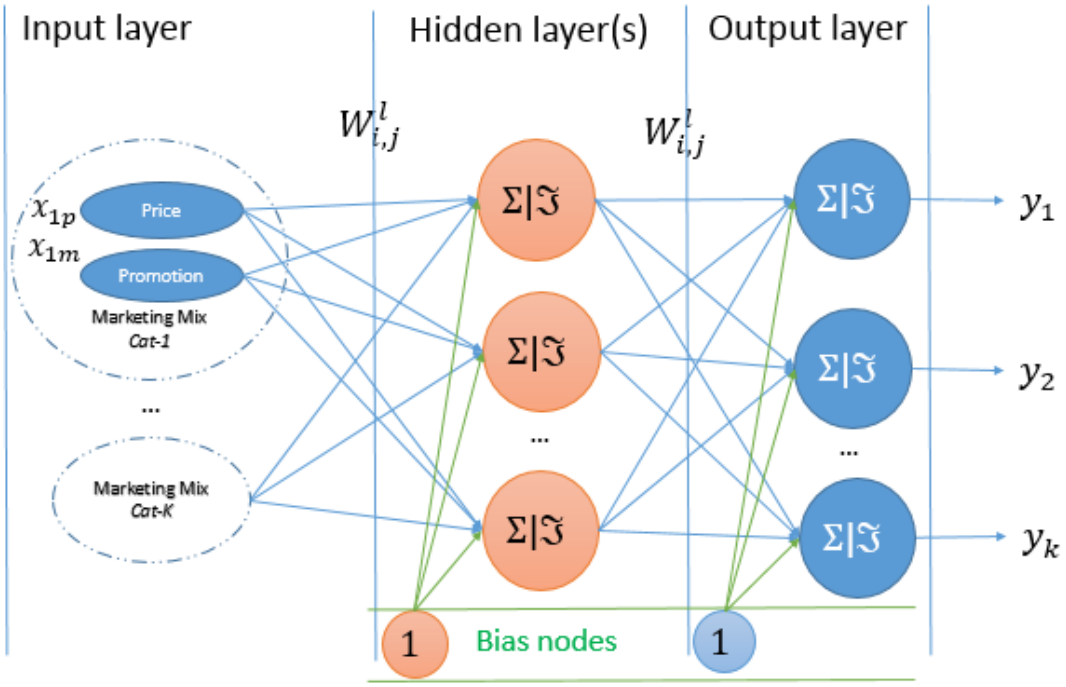


Figure 3.4 the cross category ANN model

3.2.3 Consumer Heterogeneity in ANN

ANN model places all input nodes in the same layer, namely the input layer. Within the literature reviewed, multiple layers of input nodes are currently not supported by ANN modeling. In comparison to the hierarchical MVP model, the ANN does not pre-specify hierarchical structure between IVs to DVs, but instead, learns the distance between IVs to DVs from training data. The distance is encoded in the weights parameter and in the structural connections among nodes. Appendix B provides an example to explain the mechanism.

This feature of ANN models free analysts from having to specify a hierarchical model. ANN can learn such relationships from training data, assuming that the data present enough information.

3.2.4 Configuration Tuning for Cross Effect Analysis

- Cross entropy error function

An adjusted cross entropy function, as an error function, is used in the ANN model.

Kline and Berardi (2005) reexamine the cross entropy error and mean squared error as error functions of the ANN model. Compared with previous studies, their work uses more data samples and variables. They found that cross entropy function can generate more accurate posterior probability estimation. Nielsen (2015, Chapter 3) explains a main advantages of this error function over sum of squared error function (i.e., faster learning). The function form of the cross entropy error can be expressed as equation (3.10).

$$C(y, \hat{y}) = \frac{-1}{n} \sum_{i=1}^{n} [\, (y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)\,] \qquad (3.10)$$

The y represents observed values of the dependent variable; $\hat{y}$ represent ANN estimated values of dependent variable. The right hand side sums errors of each observation and divides the sum of error by the total number of observations. Cross entropy error function is another commonly error function. Nielsen (2015, Chapter 3) provides details of this error function.

- Adjusted cross entropy cost function

The observed dependent variable is binary, either 0 or 1. So mistakes are of two types, false positive and false negative. The original cross entropy function

treats the two types of error as equally important. The error function is adjusted to make the error type of false positive less influential, as shown in equation (3.11).

$$C(y,\hat{y}) = \frac{-1}{n}\sum_{i=1}^{n}[\ (y_i * \log(\hat{y}_i) + 0.5 *\ (1 - y_i) * \log(1 - \hat{y}_i)] \qquad (3.11)$$

The adjusted function, compared with original function, retains the qualification of NN cost function such that (1) it is differentiable and monotonic (2) it is positively associated with error amount (Nielsen 2015, Chapter 3). But at the same time, it takes false positive error as less important than false negative error. This is a nice feature in marketing problems because, practically, losing a sales opportunity is much more costly than sending unsolicited mail to unknown potential customers. Technically, this adjustment actually sacrifices prediction accuracy at points of negative response, in exchange for flexibility of the ANN model to adjust its parameters for better prediction accuracy at points of positive response.

To verify the desirable feature, the ANN model is run with original cross entropy and adjusted function separately. The model's prediction performance is compared between adjusted model and base model. It is expected that the outcome of the adjusted model tends to have low error rates, as measured by Mean of Percentage Error (MPE). See Appendix C for a discussion of MPE.

Table 3.3 shows the experiment results.

Table 3.3 Effectiveness of adjusted cross entropy error

Model		y1+y2+y3+y4 ~ (p1+p2 + p3+p4 + m1+m2 + m3+m4)					
Data		Test: 4,322 X 12			Training: 17,284 X 12		
Results							
		Cutoff = 0.5			Cutoff = 0.3		
		Rep-1	Rep-2	Rep-3	Rep-1	Rep-2	Rep-3
Test data MPE	Original	178.6	329.1	196.5	29.3	275.5	164.6
	Adjusted	183.04	128	183.6	163.38	28.79	33.5
Training data MPE	Original	324.6	1172	669.3	34.8	1149	308.8
	Adjusted	623.4	116.2	625.7	36.7	22.46	33.68

As expected, the error function MPE favors the adjusted cross entropy function in 2/3 replication runs. To rule out possible noise, scores are compared in both training and testing data, and using two cutoff scores. As Table 3.3 shows, the result is consistent from training data to test data, and from cutoff score 0.5 to 0.3. Because test data are not used in training procedures, the performance on test data indicates the external validity of the trained model. Another observation is that the prediction performance of the trained NN model (with adjusted cost function) has a smaller score variance among three replications.

- Learning algorithm

The resilient back propagation learning algorithm is used in this model. Resilient back propagation (Riedmiller 1994, Günther and Fritsch 2010, Rojas 2013 page 208) is a revision of the back propagation algorithm. The traditional gradient descent algorithm uses a fixed learning rate (recall section 3.2.1). The main idea of the resilient method is to use a dynamic learning rate based on recent learning speed. This method tends to have better learning performance (Günther and Fritsch 2010).

- Activation function: logistic function output range 0 to 1 that fits the data

The observed response variable y is coded as 0 or 1, purchased or not purchased. Logistic (sigmoid) function fits this coding schema because the logistic function, expressed as $f(x) = \frac{1}{1+e^{-x\beta}}$, has the output range of 0 to 1.

- L2 (weight decay) Regularization:

This is a commonly used technique to control for overfitting. The method described in (Nielson 2015, chapter 3) is used to setup the decay parameter based on training sample size.

- Choosing number of hidden nodes

The available methods of choosing the number of hidden nodes are mainly experimental (Lendaris et al. 1993, Bishop 1995 page 170). This question is related to the theory of local/national minima of a multilayer network function (recall Bishop 1995, page 170 for details). Research on this subject is out of the scope of this project. Recent findings of (Baczyński and Parol 2004) are used as the rule. They suggest that number of weights should be one order of magnitude less than the number of learning facts in order to avoid high risk of over-fitting. Although their finding mainly addresses over-fitting on number of hidden nodes, it is assumed that when model over-fitting is to be controlled, the mode's fitting capability is guaranteed.

Specifying an over-capable ANN model (having number of hidden nodes more than necessary) not only sets the learning procedure at risk of model over-

fitting, but also can tremendously increase the requirement of computation power. Thus, the rule of (Baczyński and Parol 2004) provides a basic idea of choosing optimal number of hidden nodes.

To provide more insights on this matter, this study compares the prediction improvement from having 12 hidden nodes to having 16 hidden nodes on 16 categories model, and from having 12 hidden nodes to having 32 hidden nodes on 32 categories model. Moreover, prediction outcome is also compared between one hidden layer and two hidden layers. The outcome is reported in a later chapter.

3.3 Data

This study uses the database *Nielsen Consumer Panel and Scanner*[2] databases focused on archival data of year 2010 and 11 US cities. Information on availability and access to the dataset is available at http://research.chicagobooth.edu/nielsen.

3.3.1 Four Levels of Category Scale

To test model performance in increasing category scale, a dataset is extracted containing a set of categories that are derived from four base categories. The four base categories, *cake mix, frosting, detergent* and *softener,* are selected because they are commonly used in cross effect analysis, such as (Manchanda et al. 1999, Duvvuri et al. 2007). For each base category, the model goes upward to

[2] Calculated (or Derived) based on data from The Nielsen Company (US), LLC and provided by the Marketing Data Center at The University of Chicago Booth School of Business.

its corresponding category group and selects all categories for the category group. This procedure yields 64 categories. After removing missing data and dropping infrequent categories (purchased less than 52 times in a year), a database with 32 categories remains.

Four levels of category scale are created. The first level is the base category set. Then four more categories are added by randomly choosing one sibling category for each base category. This give us the second level of eight categories set. Then, this procedure is repeated to form the 16 and 32 categories sets. These four levels of category scale provide an approximation to the increasing category scale. These four levels are used to conduct experiments on model performance of both the MVP and the ANN model.

3.3.2 Steps and Conditions Used in Data Extraction

Households meeting following conditions are selected into data sample, (1) Nielsen Designated Marketing Areas (DMA code) in one of the eleven cities, (2) made at least two purchases of each category in ten categories of the total 32 categories.

Trips and purchases are extracted corresponding to the extracted households. An observation is a shopping trip consisting of price and promotion information of each of the 32 categories, as well as purchase decision on each of those categories. The household characteristics are attached to each observation.

3.3.3 Resulting Data Statistics

There are missing data caused by missing information of price or promotion. As a result, when there are more categories loaded into the model, there are more chances of missing information. The statistics of valid trip information is shown in Table 3.4 below.

Table 3.4 Data extraction information

Period	2010		
DMA area (11 cities)	New York, Chicago, St Louis, Dallas, Los Angeles, Boston, Houston, San Francisco, Seattle, Atlanta, Minneapolis		
Total Number of Categories	32		
Trip information			
Scale of categories	Valid trips	Valid households	Trips/household
4	21,606	1,705	12.67
8	19,399	1,611	12.04
16	19,182	1,603	11.97
32	18,058	1,566	11.53
Cause of data missing	Unavailable price and promotion for a specific category and a specific store.		
Reasonable Trips/household	Trips of non-purchase of listed categories are dropped. On average, household of our dataset make purchase once a month for the listed categories.		

Table 3.5 shows information of the 32 categories that are included in our model.

Table 3.5 Category List (highlighted are the four base categories)

ID	Module code	Module description	Number of Purchase
1	1375	MIXES - CAKE/LAYER - OVER 10 OZ.	2,451
2	1372	FROSTING READY-TO-SPREAD	1,685
3	7012	DETERGENTS - HEAVY DUTY - LIQUID	3,147
4	7060	FABRIC SOFTENERS-LIQUID	1,109
5	1343	BREADING PRODUCTS	1,049
6	1350	CROUTONS	855
7	1358	PIE & PASTRY SHELLS-PREPARED	288
8	1364	STUFFING PRODUCTS	1,008
9	1380	MIXES - BROWNIES	1,855
10	1381	MIXES-MUFFIN	1,340
11	1383	MIXES-BREAD	692
12	1384	MIXES-DESSERT-MISC.	217
13	1386	MIXES-ROLLS & BISCUITS	478
14	1387	MIXES-COOKIE	880
15	1389	CAKE DECORATIONS & ICING	378
16	1395	MIXES - PANCAKE	934
17	1396	YEAST - DRY	307
18	1435	COCONUT	79
19	1436	BAKING CHOCOLATE	326
20	1437	CHOCOLATE CHIPS & MORSELS	1,009
21	1469	BAKING POWDER	138
22	1470	BAKING SODA	305
23	7003	DETERGENTS-PACKAGED	362
24	7008	DETERGENTS - LIGHT DUTY	1,759
25	7015	LAUNDRY TREATMENT AIDS	501
26	7020	AUTOMATIC DISHWASHER COMPOUNDS	1,264
27	7025	DISHWASHER RINSING AIDS	152
28	7041	DETERGENT BOOSTERS	402
29	7062	FABRIC SOFTENERS-DRY	491
30	7080	BLEACH - LIQUID/GEL	627
31	7176	SPOT & STAIN REMOVERS	257
32	7850	LAUNDRY & IRONING ACCESSORIES	168

Table 3.6 shows a portion of the joint-purchase statistics (Appendix A.2 shows the complete table). Highlighted are values larger than 99. The highlighted

parts show that a category may be co-purchased with many other categories. For example, category y1 is most frequently purchased with y2, but many consumers purchase it with y9.

Table 3.6 Part of pair-wise Joint purchase frequency

* Cells value larger than 99 are highlighted

* Read from row perspective:
row y1 in column y1 is 2451, the total number of trips that category 1 is purchased
row y1 in column y2 is 963, the total number of trips that both category 1 and 2 are purchased
Thus, 963 is part of 2451.

	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11	y12	y13	y14	y15
y1	2451	963	176	56	69	40	30	74	258	121	67	18	51	122	93
y2		1685	118	45	52	28	20	52	228	86	48	14	33	137	94
y3			3147	337	90	61	16	76	118	115	49	9	31	64	17
y4				1109	40	20	17	28	54	35	22	5	14	25	8
y5					1049	35	9	61	46	50	25	8	22	21	9
y6						855	6	28	49	25	22	3	11	22	4
y7							288	21	20	14	12	8	7	15	2
y8								1008	43	50	27	7	7	39	12
y9									1855	89	78	29	37	140	19
y10										1340	86	16	26	53	17
y11											692	11	22	63	5
y12												217	4	14	1

3.4 Study – 1 Large-Scale MVP and Spending Habit Heterogeneity

The first study fits the MVP model (section 3.1) to the extracted dataset. This study examines performance of computation time, parameter estimation reliability, and prediction accuracy under increasing scale of categories.

3.4.1 Experiments to Study Model Performance with Increasing Scales

By replicating the MCMC (recall Section 3.1) run 10 times for each category scale, a series of performance indicators are measured and examined. Computation time is directly measured as the duration from procedure start to end.

Reliability of a parameter estimation is measured by the variance of the parameter's estimate. Variance in parameter estimations arise because of the simulation nature of the estimation method. The parameter estimation of the 4 base categories under 4, 8, 16 and 32 category scales are tracked. Comparing estimation over an increasing scale sheds light on the issue that estimations are biased under small scale (Chib et al. 2002, Boztuğ and Reutterer 2008). For a category in a scale set, its cross effect from any other category is estimated. Then, its paired partner is determined by the largest size of estimated cross effect. For example, assume that the cross effects from categories B, C, D to A are B->A = 0.2, C->A = 0.5, and D->A = 0.7. Then A is paired with D as cross effect partners. Even though prior assumptions predicts partners for the 4 base categories, the model allows partner shifting which means that a category's cross effect partner is different between a small scale model and a large-scale model. A shifting implies discovery of new knowledge of additional or unfamiliar complementary/substitution consumption in different model scales.

Prediction accuracy is measured by extending a prediction hit rate as defined by (Manchanda et al. 1999). This measure is discussed in the next section. For a given dataset, training data and test data are prepared by taking 85% and 15% random samples. The model is fitted to the training data, and predictions are made with the test data. Model fitting and predictions are repeated 10 times for each of the 4 scale datasets (whenever the computation can be finished in a reasonable time period). Model prediction accuracy is assessed by the mean and variance of prediction hit rates over the 10 replications.

3.4.2 Prediction Hit Rate, and Measures Used in This Study

- Prediction hit rate

Manchanda et al. (1999) introduce the prediction hit rate measure that is for prediction accuracy of a market basket composition. The formula is

$$1 - \frac{\sum_i abs(t_i - p_i)}{total\ trips} \tag{3.12}$$

To explain this measure, suppose we have a market basket profile as shown in the Table 3.7 below.

Table 3.7 Example of market basket composition and predictions

$\boldsymbol{i}$ Basket index	$\boldsymbol{t_i}$ True number of purchases	$\boldsymbol{p_i}$ Predicted number of purchases	Manipulated Worst prediction	Manipulated 0 prediction
1	10	12	0	20
2	20	14	0	5
3	5	9	35	0
Total	35 $\boldsymbol{total\ trips}$	Hit rate = 0.66	Hit rate = $\boldsymbol{-0.71}$	Hit rate = 0

The first column is the index of baskets. In this example, there are three types of baskets. The second column is the true number of purchases of each basket. The third column is the predicted number of purchases of each basket. The fourth and fifth columns are for the purpose of demonstrating the hit rate value range.

As shown in the third column, applying formula (3.12) gets the result:

$$1 - \frac{\sum_i [abs(10-12), abs(20-14), abs(5-9)]}{total\ trips = 35} = 1 - 0.34 = 0.66$$

Manchanda et al. (1999) suggest that the value of this hit rate measure ranges from 0 (zero prediction) to 1 (perfect prediction).

It can be found that this measure theoretically ranges from -1 to 1. As demonstrated in columns four and five, one can manipulate the prediction composition to make the formula (3.1) output zero or a negative value greater than -1. As the fourth column shows, the rule to output the worst hit rate is to predict the total trips on the least purchased basket and predict zero on each of the other baskets. This results in a hit rate of the range of -1 to zero. If the lowest true purchase is zero, then this results in a hit rate of -1.

The negative hit rate can be avoided when zero is not allowed in columns of true purchases and predictions. However, when the model scale increases, the prediction of zero becomes unavoidable. Thus, the negative hit rate is highly

possible when the model scale becomes large. We observe the negative hit rate in the scale 32 ANN model.

It is worthy to note that even though this measure can result in a negative value, a larger value still indicates better prediction accuracy.

A major issue is that the prediction hit rate is not a fair measure when it is used to compare prediction accuracy between different model scales. The problem and solution are elaborated in the next section.

- Hit Rate lift

The prediction hit rate is not a fair measure for comparing prediction accuracies between models of different scales. A larger scale model has a prediction space of much more dimensions, i.e., the number of baskets to predict purchases on[3]. A larger scale model tends to have more chances to make mistakes in predictions as demonstrated below.

Table 3.8 Hit rate measure on scale 2 and scale 4 model outcomes

	Scale 2 true purchases	Naive Prediction	Scale 4 true purchases	Naive Prediction
Basket 1	15	10	12	5
Basket 2	5	10	5	5
Basket 3			2	5
Basket 4			1	5
Total purchases	20		20	
Total error		10		14
Prediction hit rate		0.5		0.3

[3] A scale 4 model (4 categories) could include 16 possible baskets because it is the combination of 4 binary variables ($\mathbf{2^4}$). A scale 8 model could include 256 possible baskets ($\mathbf{2^8}$).

Table 3.8 shown above exemplifies the unfairness of using the hit rate to compare prediction accuracy between different model scales. The columns two and three show the scale 2 model. Given the true purchases, a naïve prediction (simply predicting the average for each basket) can reach a hit rate of 0.5. But for the scale 4 model, a similar approach of naïve prediction can only achieve a hit rate of 0.3. Comparing hit rate 0.5 with 0.3 and claiming a better prediction performance of the scale 2 model than that of the scale 4 model is not fair because both the predictions are naïve.

A fair measure can use the naïve prediction performance as a baseline.

Because the naïve prediction can always be made and achieve a prediction hit rate regardless of model scales. It can be used as the baseline performance associated with a specific scale. Follow this logic, the last row of column three and five shows that the scale 2 model has a baseline score of 0.5 and the scale 4 model has that of 0.3. The baseline score indicates the naïve prediction hit rate associated with a specific model scale and dataset.

A prediction's performance is then adjusted as the relative value to the baseline score (i.e., the hit rate lift). The formula is:

$$Hit\ rate\ lift\ =\ calculated\ hit\ rate\ -\ baseline\ hit\ rate$$ [4]

[4] The alternative measure, $\frac{calculated\ hit\ rate}{baseline\ hit\ rate}$, is subject to the bias of small baseline hit rate. The calculated lift rate would be extremely exaggerated when the baseline hit rate is close to zero. Additionally, the calculated lift rate would be negative if the baseline hit rate is negative.

The hit rate lift reflects that, comparing with the baseline score, how much the prediction hit rate is improved. The value range is -2 to $+2$, and larger values indicate larger improvements from its baseline score. Negative values indicate that a prediction performance lower than baseline score (it does not beat the naïve prediction). The hit rate lift is a fairer measure for comparison between different model scales. An example is shown in the table below.

Table 3.9 Demonstration of hit rate lift

Scale 2 true purchases	Naïve (Baseline) prediction	Models' prediction	Scale 4 true purchases	Naïve (Baseline) prediction	Models' prediction
15	10	16	12	5	11
5	10	4	5	5	4
			2	5	3
			1	5	2
Total error	10	2		14	4
Hit rate	0.5	0.9		0.3	0.8
Hit rate lift		0.4			0.5

Table 3.9 extends the Table 3.8 and exemplifies the calculation of hit rate lift (columns three and six). Column three (column six) shows a prediction of the scale 2 (scale 4) model, its hit rate, and the hit rate lift. The scale 2 hit rate, 0.9, is better than the scale 4 hit rate, 0.8. But the hit rate lift is better in scale 4 than in scale 2, as 0.5 over 0.4. Comparing the lift scores indicates that the prediction in scale 4 model is able to beat the naïve prediction more than the prediction in the scale 2 model. Thus, it has a better prediction capability.

- Base 4 categories hit rate

The prediction performance for only the base 4 categories can be calculated in any larger scale models. The example shown below in Table 3.10 demonstrates the calculation method that aggregate purchase numbers of the scale 8 model into

purchases of the base 4 categories. Using this method, this study reports the base 4 category hit rate in larger model scales. This measure provides information about whether the prediction performance is improved by using more categories (more information).

Table 3.10 Calculation of base 4 categories from a scale 8 model

Basket composition of scale 4 (c1,c2,c3,c4)	Scale 4 purchases (aggregate by the base 4 categories)	Corresponding composition of scale 8	Scale 8 purchases
1 (0,0,0,0)	30	(0,0,0,0, 0,0,0,1)	10
		(0,0,0,0, 0,0,1,0)	10
		(0,0,0,0, 0,0,1,1)	5
		(0,0,0,0, …)	5
2 (0,0,0,1)	20	(0,0,0,1, 0,0,0,0)	3
		(0,0,0,1, 0,0,0,1)	7
		(0,0,0,1, …)	10
3 (0,0,1,1)	30	(0,0,1,1, 0,0,0,0)	17
		(0,0,0,1, …)	13
4 (0,1,1,1)	…	…	…
….	…	…	…

3.4.3 Parameter Estimation

The model specified in section 3.1 is implemented by customizing an open source *R* package *bayesm* (Rossi 2012). *R* is an open source data analysis tool (R Core Team 2014). At this time, a hierarchical MVP model with random effect heterogeneity is not yet implemented in the *bayesm* package. The estimation program extends the function *rmvpGibbs* in the *bayesm* package. The implementation is based on the work of Allenby et al. (2005) and is referred to as the model by Duvvuri et al. (2007).

In general, the estimation procedure is a MCMC method guided by Bayesian inference statistics. Recall section 3.1.4 for a brief review of this method. Rossi et al.'s book is an helpful resource for this subject (Rossi et al. 2005).

3.5 Study – 2 Large-Scale ANN and Non-linear Effect

The ANN study trains the ANN model specified in section 3.2 with extracted datasets. Datasets used in this study are identical to those used in Study - 1. Model performance such as computation time, and prediction hit rate are examined under increasing scales of categories. Importantly, this study examines non-linear relationship between IVs and DVs (recall section 3.5.3)

The model specified in section 3.2 is implemented by customizing an open source *R* package *neuralnet* (Günther and Fritsch 2010). *R* is an open source data analysis tool (R Core Team 2014).

3.5.1 Experiments to Study Model Performance with Increasing Scales

To examine model computation time and prediction hit rate, this study uses a procedure that is exactly the same as that of Study – 1 (recall section 3.4.1). The exception is that estimation reliability is not available in this ANN study (recall section 3.5.2).

3.5.2 Nonparametric Model ANN

ANN models are nonparametric models (Intrator and Intrator 2001, Rasmussen and Williams 2006). Thus, the reliability of parameter estimations is

not a meaningful concept in the ANN study. Regression models output the beta coefficient (i.e., the estimation of the general effect from an IV to a DV). It can be expected that a corresponding estimation in the ANN model occurs. However, ANN models are stochastic processes in which the resulting solutions are random (Rasmussen and Williams 2006). In other words, stochastic processes do not have a theoretically unique solution. For example, suppose that ten replication runs for an ANN model output ten resulting solutions and the solutions are very different from each other. Even though the ten solutions are very different, it is not meaningful to claim that the parameter estimation is unreliable, because any of the ten solutions are a correct solution to the ANN problem.

Even though the ANN technique does not provide general effect parameters as regression models do, the ANN technique is useful for a broad range of applications for which prediction is the man purpose. Examples include, but are not limited to, image recognition and natural language processing.

3.5.3 The Special Feature Reported by ANN Model

One major advantage of ANN over MVP is the ability to capture non-linear relationships. The relationships are learned and embedded in ANN's structure. For example, a linear effect of price change on purchase intent could be that, on average, one dollar increase of price is associated with 50 percent decrease of purchase probability. A non-linear effect could be that, when the price is 20 dollars, one dollar increase of price is associated with a 50 percent decrease of purchase probability; when the price is 30, the same price increase is associated with only

10 percent decrease; when the price is 40, it is associated with a negligible decrease.

From the example above, an IV may have different regions where the effect size and variance may be different. Such regional effects reflect a non-linear relationship. To help identify such non-linear relationships, Intrator and Intrator (2001) introduce the approach of plotting generalized weights which plots calculated generalized weights value over the observed values of an IV. Such a plot shows information of an IV's regions where the IV has non-linear impact on a DV. Managers could use such information, because they can set the value of an IV to the desirable positons to optimize their results.

Chapter 4. Results of Study One

This chapter implements Study – 1 and reports findings.

4.1 Model Runs

Table 4.1 shows several parameter choices of our MCMC simulation, as well as the reasons why these values were chosen.

Table 4.1 MCMC parameters

MCMC Parameter	Value	Rationale
Gibbs rounds	20,000	The value is determined by experiments results. Experiments for 5000, 10,000, 20,000, 30,000 and 50,000 rounds are conducted. 20,000 rounds result in good enough converge. (Appendix A.3 shows part of plots price parameter draws)
Thinning interval	Pick every 20^{th} of draws	The rule is balance between computation time and goodness of randomness. The larger this number, the closer the MCMC draws to randomness; but the larger this number, the longer the computing time needed. Manchanda et al. (1999) uses 5th.
Burn-in rounds	10000	Manchanda et al. (1999) choose 45,000, 90% of all draws. Duvvuri et al. (2007) choose 12,500, 25% of all draws. The experiments show that 10,000 works well in this study. Appendix A.3 plots draws of price effect of the 4 base categories model. In the plotted figure, converge presents around or before 10,000 draws.

The 4-category model runs ten times on a Windows PC with Inter i7 4510u CPU (2 core, 2.6 GHz) and 16GB internal memory. The 8-category model runs eight times on the same PC and two times on a clustered dedicated server[5] with two virtual core CPU and four GBs internal memory. The 16 and 32 category models are tried on both machines. Using a personal PC to run the model is to ensure that the model is computable on a modern PC. The parameter estimation

[5] The server cluster is provided by http://aggregate.org/KAOS. Its usage was sponsored by Dr. Goldsmith of the Computer Science Department of University of Kentucky.

had similar results from Windows and Linux runs. To make the running time compatible, time to finish computing on Windows has been translated into that on Linux by matching the relative time needed to finish each single step.

4.2 Model Performance of Increasing Scale

Table 4.2 shows the summary of performance over 10 replication runs of the MVP model. The first two columns respectively show the number of categories loaded in a model and the number of parameters to be estimated. When the category number doubles the number of parameters increases by 4 times, when considering both price and promotion as cross effect IVs. The 3rd column shows the average computing resource consumption reported by the Windows Task Manager when the estimation procedure is running. The 4th column shows both mean and standard deviation of time (minutes) to finish an estimation based on 10 replications each. The 5th column shows both mean and standard deviation of prediction accuracy of the estimated model.

Table 4.2 General model performance with increasing category scale

Scale	Number of parameters	Resource Usage	Time to compute in minutes [a]	Prediction hit rate (recall section 3.4.2)		
				Prediction hit rate [a]	Hit rate lift	Base 4 categories hit rate [a]
4	4 categories * (4*p + 4*m) = 32 = 2^5	2GB Ram, 15% CPU	83.87 (0.95)	0.83 (0.0007)	1.15	-
8	8 categories * (8*p + 8*m) = 128 = 2^7	4GB Ram, 20% CPU	849.54 (136.28)	0.79(0.001)	1.30	0.863 (0.001)
16	16 categories * (16*p + 16*m) = 512 = 2^9	10GB Ram, 30% CPU	19320+ (322+ hours)	NA	NA	NA
32	32 categories * (32*p + 32*m) = 2048 = 2^{11}	Hit error of Out of Memory (trying to allocate 16GB memory)	NA	NA	NA	NA

a: values in this column are shown as mean (standard deviation) whenever applicable.

On average, the scale 4 model takes about one hour and 24 minutes to finish computing with standard deviation of about one minute. The scale 8 model takes about 14 hours and ten minutes with standard deviation of two hours and 26 minutes. The scale 16 model is estimated to require more than 322 hours to finish computing. The memory use of the estimation procedure is not constrained. In such a case, it will hit an out-of-memory error when the available memory is not ideal. Whenever the out-of-memory error does not occur, it indicates that the estimation runs with enough resources. The out-of-memory error is reached when running scale 32 model. Thus, the scale 32, model is not computable with practical resources. Category scale 4, 8 and 16 are all computable. However, the scale 16 will take too long to finish (about a week), so it is treated as not feasible. When the number of categories doubles from four to eight, computing time increases about

ten times; from eight to 16, it increases about 22 times. In general, the resources and computing time are exponentially increasing with category scales, and scale 16 model takes too long to finish computing with a modern (2015) personal computer.

The last three columns of Table 4.2 show the model prediction accuracy measured by the three types of hit rates (recall section 3.4.2). Generally, when the prediction spaces increase from 16 (2^4 for scale 4) possible baskets to 256 (2^8 for scale 8) possible baskets, the prediction hit rate does not dramatically deteriorate. The hit rate lift of 1.30 in the scale 8 model is better than that of 1.15 in the scale 4 model. It suggests higher capability of the scale 8 model because it improves more from its naïve prediction. Additionally, if considering only the base 4 categories, the prediction hit rate is improved from 0.83 (the scale 4 model) to 0.86 (the scale 8 model). It suggests that scale 8 model can provide more information to the estimation and improves the prediction hit rates on the base 4 categories.

Table 4.3 Own/Cross Effect estimation (average of 10 runs)

Id	Category	Sc-ale	Intercept	OP	OM	CPi	CP	CMi	CM
1	mixes - cake/layer - over 10 oz.	4	-1.307[a] (0.211[b]) 3[c]	-26.41 (3.449) 3	1.799 (0.344) 3	4	9.122 (1.601) 3	3	-1.329 (0.344) 3
		8	-1.001 (0.243) 3	-36.563 (4.686) 3	1.966 (0.416) 3	4	7.284 (1.706) 3	7	-1.643 (1.044) 1.7
2	frosting ready-to-spread	4	-1.877 (0.338) 3	-4.815 (1.479) 3	4.358 (0.812) 3	1	-17.678 (2.461) 3	4	-1.447 (0.678) 2.6
		8	-2.096 (0.445) 3	-15.975 (2.663) 3	5.02 (1) 3	1	-23.09 (3.352) 3	5.2	1.489 (1.358) 1
3	detergents - heavy duty - liquid	4	-0.258 (0.054) 3	-1.766 (0.236) 3	1.138 (0.155) 3	2	-1.915 (0.408) 3	2	-0.346 (0.174) 2.2
		8	-0.392 (0.068) 3	-3.003 (0.277) 3	1.239 (0.134) 3	1	1.746 (0.422) 3	7	-0.649 (0.325) 2
4	fabric softeners-liquid	4	-0.15 (0.172) 0	-14.858 (2.485) 3	3.273 (0.601) 3	2	-6.064 (1.641) 3	1	-1.635 (0.619) 3
		8	-1.738 (0.426) 3	-18.909 (3.221) 3	4.706 (0.891) 3	2	-8.169 (2.628) 3	1	-2.184 (0.924) 3
5	breading products	8	-4.709 (0.927) 3	-1.211 (0.579) 2.5	8.398 (2.014) 3	1	9.787 (2.974) 3	1	-1.363 (1.042) 1
6	croutons	8	-0.378 (0.335) 0	-16.582 (2.668) 3	2.419 (0.773) 3	2	11.213 (2.738) 3	3	-1.339 (0.619) 2.8
7	pie & pastry shells-prepared	8	-2.669 (0.736) 3	-5.26 (1.385) 3	3.33 (1.434) 3	4	10.335 (3.527) 3	8	1.331 (0.935) 1
8	stuffing products	8	-3.642 (0.969) 3	-3.735 (1.237) 3	4.086 (1.189) 3	1	8.047 (2.9) 3	4	-1.908 (1.123) 2

a: average parameter value of 10 runs
b: average standard deviation the parameter value of 10 runs
c: average significance level of 10 runs.
0-> not significant at alpha = 0.1 level
1-> significant at alpha = 0.1 level
2-> significant at alpha = 0.05 level
3-> significant at alpha = 0.01 level. Highlighted as bold is not significant at 0.01 level.
OP: own price effect
OM: own promotion effect
CPi: corresponding category of largest cross price effect
CP: cross price effect from CPi
CMi: corresponding category of largest cross promotion effect
CM: cross promotion effect from CMi

Table 4.3 shows the parameter estimation as an average of the 10 replication runs. First, the estimation has face validity in that the effect of a category's own price on latent utility is negative, and the effect of its own promotion is positive, for all categories (column OP and OM). This is consistent with the

traditional wisdom that price drop and/or product promotion increases probability of purchase.

For every category, the price effect (column OP) is consistently larger for scale 8 than for scale 4. This result suggests that consumers of 8 categories baskets are more sensitive to price change than those of 4 categories. If a manager makes predictions of category 8 customers using category 4's models, the prediction would be inaccurate.

The results support the objective of this study. It shows that relaxing prior assumptions of pairing categories for cross effect analysis leads to discovery of new knowledge. The frosting takes the cake mix as the cross price partner, but the mix takes softener as partner. This finding indicates that the prior assumptions of cross effect is partially correct. Demand on frosting is sensitive to price change of cake mix, but demand of cake mix is more sensitive to price change of softener. Apparently, there is a direction for cross effect.

Comparing OP with CP, the OP of cake mix remains a higher value for both scale 4 and scale 8, but the CP has lower value for both scales. This finding suggests that the demand of cake mix is mainly influenced by its own price, but less influenced by the cross price effect. So its demand is more autonomous. In contrast, the OP of frosting remains at a lower value and CP has higher value for both scale 4 and scale 8. This suggests that frosting is a complementary category which means that its utility is influenced more by its cross price, rather than by its own price. Using this approach, the result indicates that the softener is an

autonomous category because its OP is larger than CP. Detergent is a semi-complementary category because its OP and CP are at a similar level. This finding is consistent with that of (Duvvuri et al. 2007) in that consumers can be sensitive to price change in one category, but at the same time not sensitive to that of the complementary category.

Using this methodology managers will be able to identify such *complementary* categories as well as *autonomous* categories. It is more effective to promote the autonomous categories because their price not only influence its own demand, but also influence its complementary category's demand.

Another observation is that the impact size of frosting OP tremendously increases from 4.8 to 15.9, an increase of more than 300%, when the model scale increased from 4 to 8. In contrast, the cake mix OP is -26.4 in scale 4 and -36.5 in scale 8, the impact size increases less than 100%. The change is about 1/3 of the former. An implication is that when managers want to do promotion on frosting, they could refer to a scale 4 or a scale 8 model. If the consumer is a scale 4 consumer, a higher level of promotion is needed in order to have the same impact as that to scale 8 customers. Another takeaway is that frosting is the only category that has this large increase of price impact size. Combining the finding of autonomous and complementary categories, this finding may suggest that complementary categories tend to have highly different price impact between the scale 4 and scale 8 customers.

The cross category partner remains the same from scale 4 to scale 8 for the base categories except for category 3. Detergent shifts its partner from frosting to cake mix, and its CP changed from -1.9 to $+1.7$. It suggests that, in terms of cross price effect, frosting is the influencer to detergent in scale 4 model. But in the scale 8 model that takes into account categories 5, 6, 7 and 8, cake mix would be the influencer to detergent. Understanding the different influence structure helps managers better identify consumers' motivations in scale 4 and scale 8.

Table 4.4 Parameter estimation dispersion of 4 categories and 8 categories model

Category	Scale	Average CV of 10 runs				
		Intercept	OP	OM	CP	CM
mixes - cake/layer - over 10 oz.	4	0.16	0.13	0.19	0.18	0.26
	8	0.24	0.13	0.21	0.23	NA[a]
frosting ready-to-spread	4	0.18	0.31	0.19	0.14	0.47
	8	0.21	0.17	0.20	0.15	NA
detergents - heavy duty - liquid	4	0.21	0.13	0.14	0.21	0.50
	8	0.17	0.09	0.11	0.24	0.50
fabric softeners-liquid	4	NA	0.17	0.18	0.27	0.38
	8	0.25	0.18	0.20	0.33	0.43

a: NA indicates that the average level of significance over 10 runs is less than 0.05. When the parameter is not significantly different from 0, the CV become misleading[4].
OP: own price effect
OM: own promotion effect
CP: cross price effect
CM: cross promotion effect

Table 4.4 shows coefficient of variation (CV) of the major model parameters[6]. CV is a statistical measure of distribution dispersion[7]. One advantage of CV is that variables with large or small means can be compared on the dispersion of their distribution. Thus, it can be used to compare estimation

[6] Formula is $\frac{1}{10}\sum_{i=1}^{10}\frac{\sigma_i}{abs(\mu_i)}$. μ_i is a parameter, and σ_i is its standard deviation.
[7] http://www.ats.ucla.edu/stat/mult_pkg/faq/general/coefficient_of_variation.htm

reliability between scale 4 and 8. This measure has been used in business studies such as that of Shechtman et al. (2005). If CVs in scale 8 are generally larger than that in scale 4, this may suggest that scale 8 has large variance of parameter estimations. Table 4.4 shows that CV value in scale 4 is not generally smaller than that in scale 8. For example, category cake mix and softener, estimation seems more reliable for scale 4 because all CVs are smaller than or equal to CVs of scale 8. For categories frosting and detergent, scale 4 does not have consistently lower CVs.

In general, the results show that Increase of category scale does not significantly affect the reliability of parameter estimation at least in the range of scale 8.

Chapter 5. Results of Study Two

This chapter implements study – 2 and reports findings.

5.1 Model Runs

Table 5.1 shows the ANN parameters. Reasons to choose these values are discussed in Section 3.2. For a general description of ANN model, Recall section 3.2. A general ANN model can been seen in Figure 3.4 of section 3.2.2.

Table 5.1 Summary of ANN model configurations

ANN Parameter	Value	Main reason
Error function	Cross entropy function adjusted to favor over-prediction against under-prediction	(1) Faster Learning than sum of squared error (2) Fit marketing problem
Activation function	Logistic	Fit choice encoding (0, and 1)
Learning algorithm	Resilient backpropagation	Dynamic learning rate based on current learning speed
Hidden nodes	Scale 4: 4, scale 8: 8, scale 16: 12, scale 32: 12	Optimal learning capability based on findings of (Baczyński and Parol 2004)

The model is run on both a Windows PC and a Linux clustered server with the approach being the same as MVP model estimation in Chapter 4.

5.1.1 Convergence

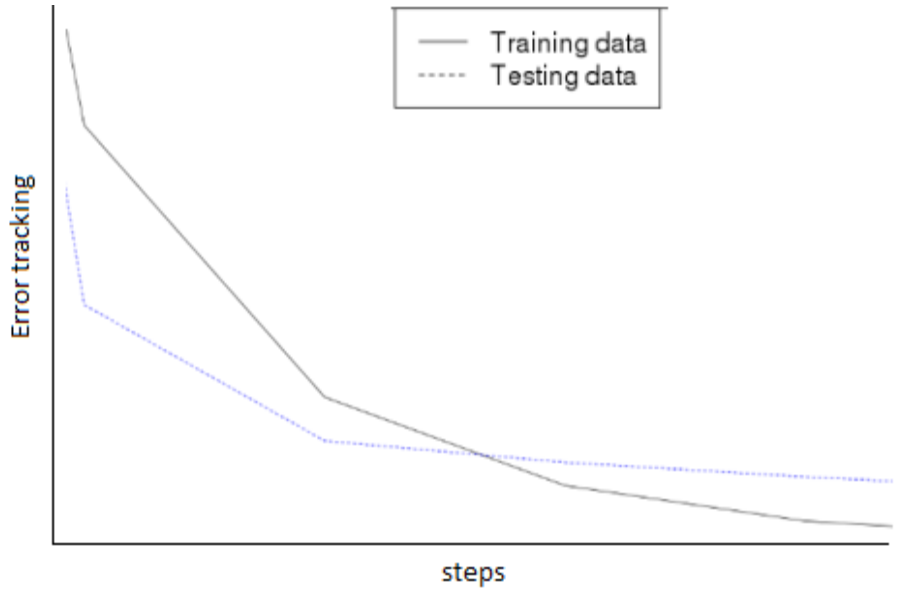
The stopping rule discussed in section 3.2.1 is to check model convergence. Another complementary technique is to monitor error tracking in the course of learning steps. Table 5.2 shows an error tracking plot of the 4 model scales. Each model scale has 10 replications, and the error tracking was plotted for one of the replications. To make it easy to read, part of the whole plot is clipped to show the turning point of convergence.

The plot tracks the error value at every 1000 steps. The error is calculated for both training data and test data. An indication of a good learning outcome is that both the training and test error go down quickly at the beginning and then slow down until training error converges. In contract, an indication of a suspicious learning outcome is anything different from a good learning pattern. Table 5.2 shows an example of good learning in scale 4, 16 and 32, but an example of suspicious learning in scale 8. Analysts have to make a decision about whether to accept or reject the suspicious learning outcome. A case of suspicious learning suggests that one or both of the training and test error function fell into a “bumpy” area where the convergence is not stable. This result suggests unstable predictions. Thus, the learned model is at high risk of non-generalizability.

The reason underlying a result of suspicious learning varies. A fact is that reduction in training error is not guaranteed at any learning step for the following reason. Each weight parameter represents a dimension of the error function. Changing a parameter value will change the location of the error function. Different locations of the error function could invert the calculated gradient descent on a weight parameter. Thus, simultaneously updating all weights, even though toward the gradient descent direction in each individual dimension, does not guarantee minimization of an error function.

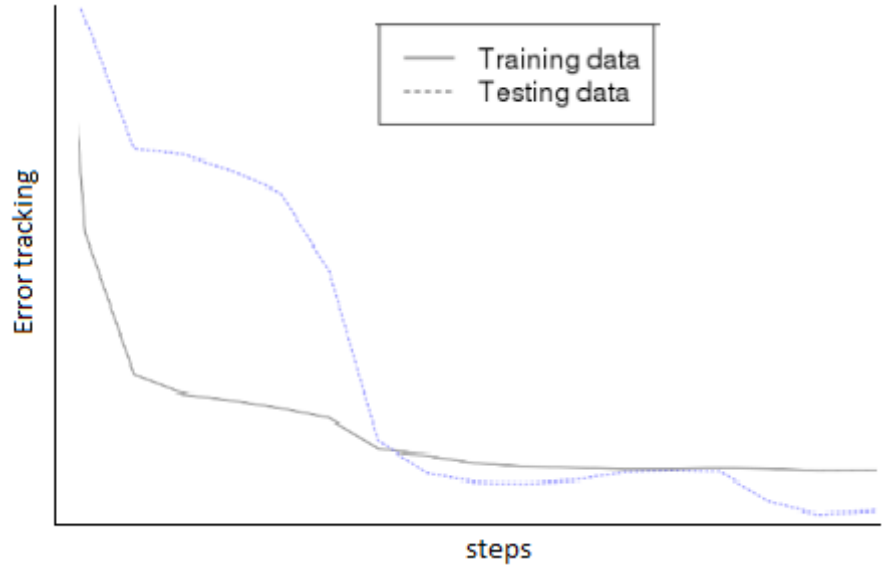
Table 5.2 Selected error tracking plot

Scale = 4

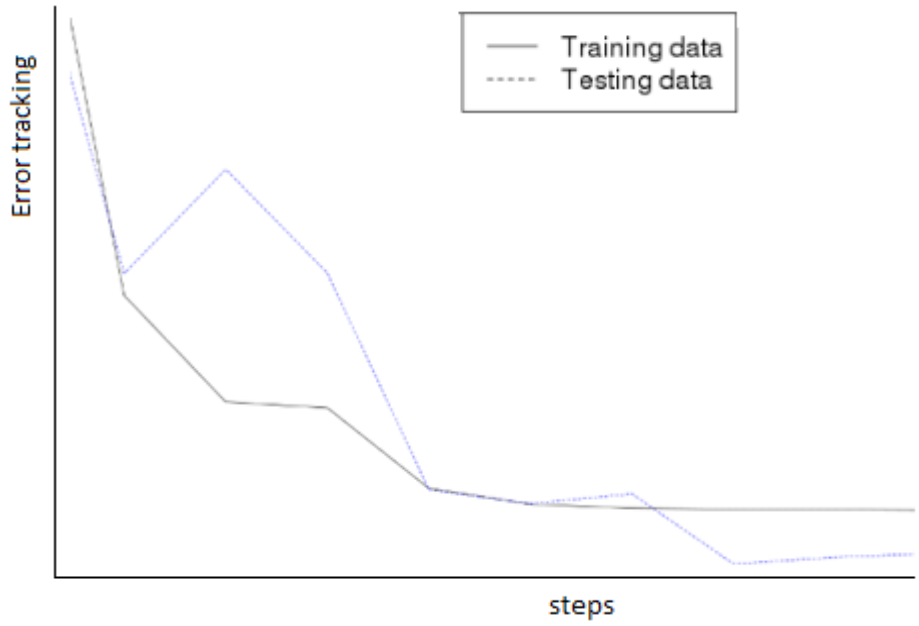


Scale=8

Scale=16



Scale=32



5.2 Model Performance with Increasing Scale

Table 5.3 shows the summary of performance over 10 replication runs of the ANN model. The first two columns, respectively, show the number of categories loaded in the model and number of parameters to be estimated. When the category number doubles, the number of parameters increases by four times when considering both price and promotion as cross effect IVs. The third column shows the average computing resource consumption reported by the Windows Task Manager when the estimation procedure is running. The fourth column shows the mean and standard deviation of time (minutes) to finish an estimation based on ten replications. Columns 5, 6, 7 and 8 show the prediction accuracy measures (recall section 3.4.2 for the discussion of these measures).

Table 5.3 General model performance with increasing category scale [(a)]

Scale	Number of parameters	Resource Usage	Time to compute [a]	Prediction hit rate (recall section 3.4.2)		
				Hit rate [a]	Hit rate lift	Base 4 categories hit rate [a]
4	(4*p + 4*m) * (4+1) hidden nodes * (4+1) output nodes = 200	0.12GB Ram, 32% CPU	2.92 (1.64)	0.82 (0.004)	1.15	-
8	(8*p + 8*m) * (8+1) hidden nodes * (8+1) output nodes = 1296	0.13GB Ram, 32% CPU	92.60 (36.46)	0.73 (0.01)	1.23	0.83 (0.05)
16	(16*p + 16*m) * (12+1) hidden nodes * (12+1) output nodes = 5408	0.16GB Ram, 32% CPU	176.27 (198.68)	0.55 (0.01)	1.04	0.88 (0.01)
32	(32*p + 32*m) * (12+1) hidden nodes * (12+1) output nodes = 10816	0.24 GB, 32% CPU	247.73 (207.90)	-0.30 (0.04)	-0.02	0.31 (0.16)
a: values in this column are shown as mean (standard deviation).						

Holding the error rate fixed at 0.05, on average, the scale 4 model takes about three minutes to finish computing with about 1.6 minutes standard deviation. The scale 8 model takes about one and half hours with standard deviation of 36 minutes. The scale 16 model takes about three hours with standard deviation of three hours and 18 minutes. The scale 32 model takes about four hours and eight minutes with standard deviation of three and half hours. When memory use is not constrained, the estimation procedure and models of all the four scales do not hit an out-of-memory error. All four scales can be computed within a few hours. In general, the resources and computing time do not exponentially increase with category scales in the current model setting.

The prediction hit rates keep decreasing. For scale 32, this measure becomes unusable. The hit rate lift shows that the rank of model capability is scale 16, 8, 4 and 32. In scale 4, 8, and 16, the larger the scale, the higher the model's prediction hit rate lift. But the scale 32 probably had a low signal to noise ratio and the ANN model is "confused" so that the outcome model cannot beat even the naïve prediction. Note that each model is run at least 10 times with random number as starting weights. Additionally, if considering only the base 4 categories, the hit rate is increasing from scale 4 to 8 and 16. This result supports the idea that including more relevant categories to model can provide useful information and improve prediction accuracy. But again, with too many (32) categories the model can be overloaded and confused.

The failure of the scale 32 model may be attributed to insufficient model capability, because the number of hidden nodes is set by a rule from the literature

(recall section 3.2.4 for the rule used). The number of hidden nodes are set to the upper limit of the rule suggested by (Baczyński and Parol 2004). The rule may not robust enough to cover a very large-scale problem like the scale 32 model in this study. To verify this issue, a test run is conducted to use 32 hidden nodes instead of the original 12 hidden nodes. This test takes about 15 hours to finish the computation and its hit rate reaches about 0.47 with lift of 0.75. This is evidence that the 12 hidden nodes ANN has insufficient capability to model the scale 32 model. Even though using 32 hidden nodes improves the model hit rate, the lift rate is still the lowest. Additionally, the base 4 category hit rate reaches 0.80 which is highly improved from the 12 hidden nodes model (0.31). But again, the score is still lower than the scale 4 model's (0.82).

5.3 Non-linear Relationship

To study non-linear relationships between IVs and DVs in the cross effect analysis context, we use the Generalized Weights concept introduced by Intrator and Intrator (2001). The formula to calculate the Generalized Weight of an IV to a DV is

$$\mathrm{w_i} = \frac{\partial \log\left(\frac{\mathrm{o(x)}}{1-\mathrm{o(x)}}\right)}{\partial \mathrm{x_i}} \tag{5.1}$$

The function $\boldsymbol{o}(\boldsymbol{x})$ indicates the output of a NN model given a set of inputs $\boldsymbol{x}$. The calculated quantity, w_i, is the first derivative of log-odds of output value over an input variable $\boldsymbol{x_i}$.

To better understand this quantity, one needs to take the approach of interpreting logistic regression[8]. The odds ratio $\frac{o(x)}{1-o(x)}$ is the ratio of the probability that an even happens over that it does not happens. For example, if $o(x)$ represents the probability that a category y1 is purchased, then an odd-ratio 1.16 means that the probability that y1 is purchased is 16% = (1.16-1) higher than the probability that it is not purchased. The logistic regression model as shown in equation (5.2) takes the logarithm of this odd-ration, called logit function, as the dependent variable. The logit function maps a probability space $o(x) \sim (0, 1)$ to the value space of $(-Inf, +Inf)$ which embodies a classic non-bounded continuous dependent variable, but at the same time remains the function's monotonicity and continuity. Specifically, the odds term $\frac{o(x)}{1-o(x)}$ transforms the value range $(0, 1)$ to $(0, +Inf)$, and the log term $\log(p)$ transforms $(0, +Inf)$ to $(-Inf, +Inf)$. In such a specification as shown in equations (5.3) and (5.4), the regression parameter b cannot be directly interpreted as the linear impact on odds, but instead the transformation e^b is.

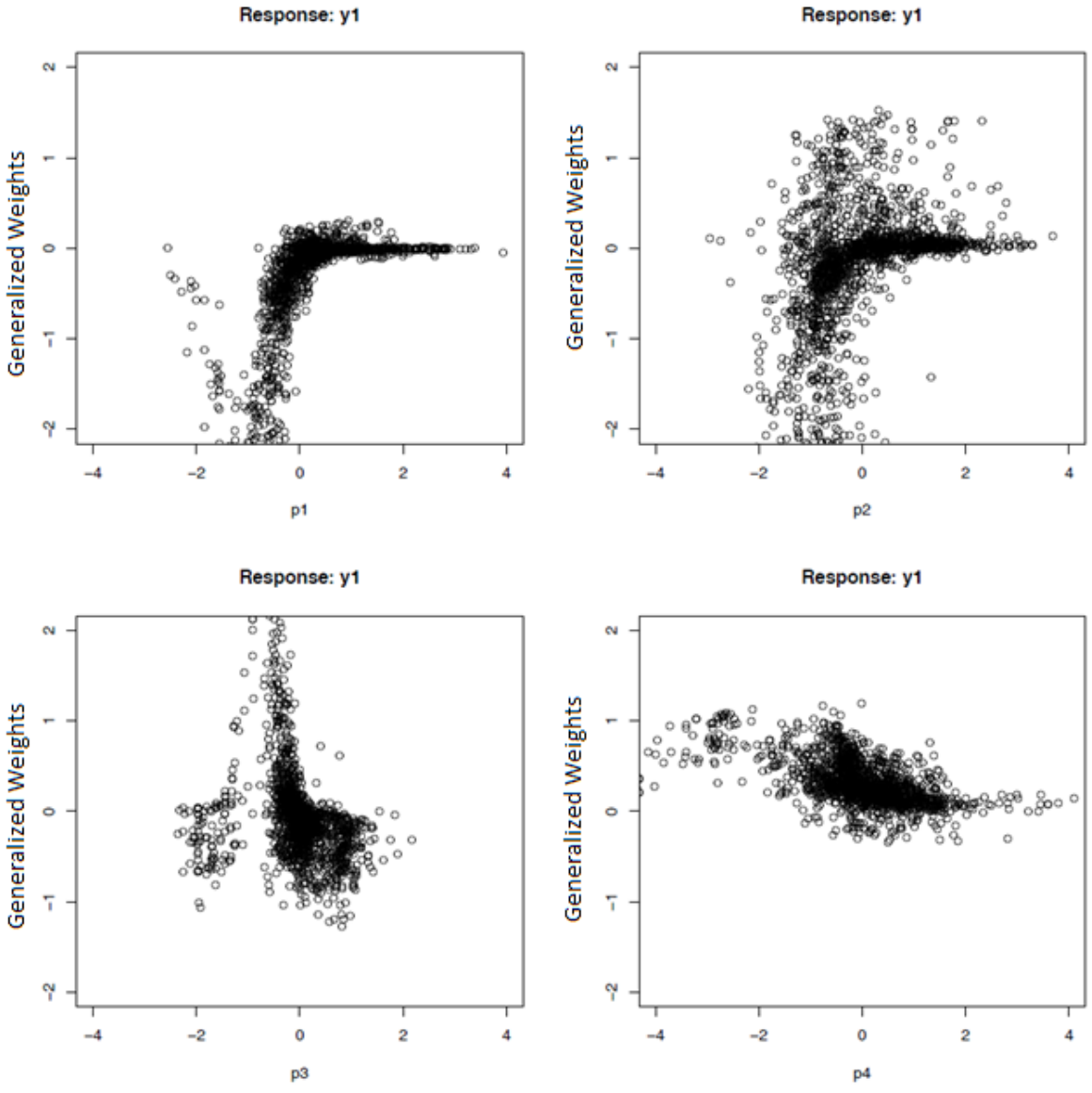
$$\mathbf{logit}\big(o(x)\big) = \log\left(\frac{o(x)}{1-o(x)}\right) = a + bx \qquad (5.2)$$

$$e^{\log\left(\frac{o(x)}{1-o(x)}\right)} = e^{a+bx} \qquad (5.3)$$

[8] Refer to Introduction to SAS. UCLA: Statistical Consulting Group, from http://www.ats.ucla.edu/stat/stata/faq/oratio.htm and http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm (accessed June 15, 2015).

$$\frac{o(x)}{1-o(x)} = e^{a+bx} \qquad (5.4)$$

With these explanations of log-odds in logit models, the generalized weights shown in equation (5.1) can be interpreted as to how the change of log-odds is associated with the change of x_i. If the x_i is linearly related to the log-odds, then this quantity of generalized weights tends to be less variant (Intrator and Intrator, 2001).



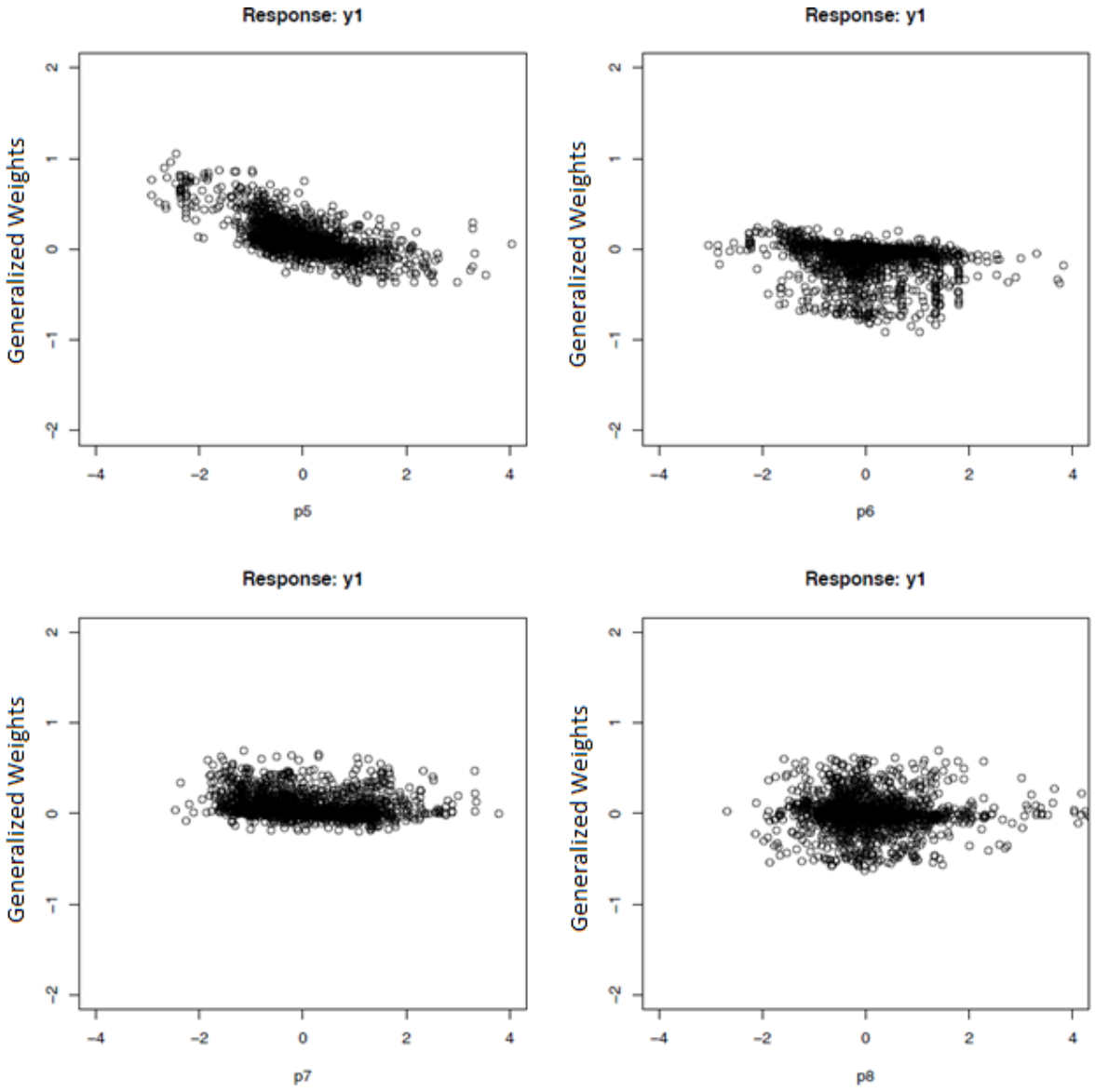


Figure 5.1 Selected plot of generalized weights.

Figure 5.1 is the plot of generalized weights from training outcome data. The X axis is the price observations for each category. The p1 indicates price of category 1, and p2 is price of category 2, and so forth. (refer to section 3.3 for category name). The Y axis is the calculated generalized weights for each data point.

Inspecting this plot reveals several relationships. First, when p1 is in the range of $(-1, 0)$, the generalized weight of p1 on y1 has the largest variance with a mean around -1. This implies that when $\mu - \sigma \leq p1 \leq \mu$, the purchase odds-ratio is not linearly related to price change. From the plot, one can see that, only in this area, when price drops the price effect on utility quickly becomes high; in contrast, when price is in the area of above average, the effect of price drop on odds-ratio does not change much. A managerial implication is that price drop will work well to largely boost sales when the current price is not higher than average price.

As a contrast, the generalized weights of p4 on y1 are mostly falling in the range of 0 to 1. It indicates that the effect is much more linear. Price drop of p4 would have a similar effect, no matter if p4 is higher or lower than its average.

These findings can be interpreted from two perspectives: the data analysis perspective and the managerial perspective. It could be the case that consumers who generate the purchases when the price is in discount range are budget buyers who are inherently sensitive to price discount. When price is high, these consumers tend to drop out of the purchase population and the insensitive buyers remains. Such a speculation can generate new research questions.

From the managerial perspective, the ANN model is able to provide generalized weights plots on all IVs to DVs. Viewing such a chart can give managers a quick view about which category is important in terms dropping price or putting out advertisements. For example, as shown in Figure 5.1, a manager

who want to boost sales of category one can decide to exclude p4 to p8 as good candidates for offering discounts, because the a discount of p1 to p3 seems more effective at attracting new purchases.

Chapter 6. Discussion

The findings of this dissertation contribute to cross effect research from three perspectives.

6.1 General Comparison between the ANN and the MVP

First, performance of existing models in the context of large-scale data is examined. Existing research for similar objectives is very limited.

Data experiments shows that the widely used MVP model becomes hard to compute when 16 categories are simultaneously estimated. In contrast, the alternative ANN model can finish computing in a reasonable time.

Figure 6.1 plots the number of model parameters and computation time over model scale for both the ANN and MVP model.

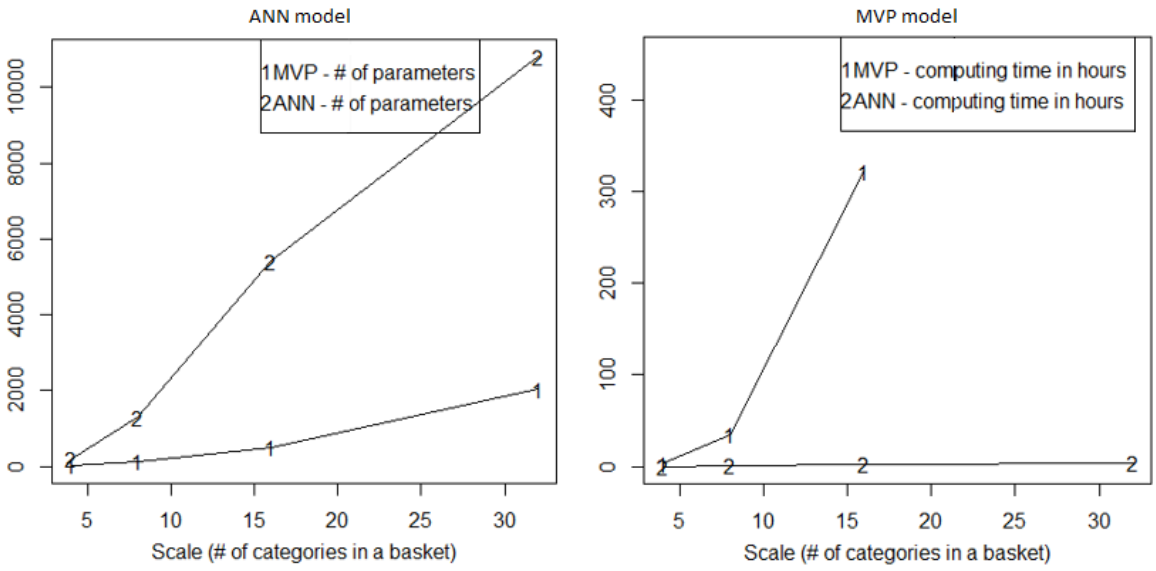


Figure 6.1 Number of parameters and computation time

The major findings of comparing the MVP and ANN model for performance are shown in Table 6.1.

Table 6.1 General Performance of the ANN and the MVP model

Performance	ANN	MVP	Discussion
Computation time	Is able to compute scale 4, 8, 16 and 32.	Is able to compute scale 4, 8. But requires more than 322 hours in scale 16, and hits out of memory error in scale 32.	To be able to utilize the calculation efficiency of vectorization, the MCMC algorithm needs to expand the data structure into a diagonal matrix form. This form exponentially increases the memory use for matrix manipulation. A modern PC with 16 GB memory hits an out-of-memory error for trying to compute a model of 32 categories.
Resource needed	Does not significantly increase with number of categories loaded	Memory requirements are doubled when the number of categories doubles.	
Parameter reliability	NA	Variance of parameter estimation remains small. It is not guaranteed that larger scale tends to have larger variance.	

In general, the ANN model is more adaptable to increasing category scales. Managers willing to utilize large-scale cross effect analysis can enjoy the scalability of the ANN model. However, resource requirements of the MVP model increase exponentially when the scale increases from 4 to 8 and16.

Table 6.2 compares mechanism of the ANN and the MVP model.

Table 6.2 Operational feasibility

Mechanism	ANN	MVP
Computation time	Analysts can adjust the result error level to balance the prediction accuracy and computation time.	Computation time is at most determined by the size of training data.
Estimation method	Both are stochastic. But the ANN has less constrains in searching for an optimal solution. The MVP relies on the assumption of normal distribution of DVs and errors.	
Operational Complexity	Compared with the ANN, the MVP needs more time and intellectual effort.	

6.2 Findings of the MVP Model Estimation

Because the large-scale cross effect model estimates cross effects between any two categories it provides an opportunity to discover evidence-based cross effect partners that may be non-intuitive. As shown in section 4.1, identified by the highest cross effect score, category frosting is paired with cake mix, but cake mix is paired with softener. If the model had only allowed the pairing between cake mix and frosting, then managers will have missed the fact that consumers' utility on cake mix is actually more influenced by the price of softener than by the price of frosting!

A Large-scale dataset includes more categories. Thus, it provides more information to the model, which is then able to identify partners miss-specified in a small scale data model. For example, the results shown in section 4.1 reveal that detergent is most influenced by the price of cake mix in the scale 8 model; but it is most influenced by frosting in the scale 4 dataset. This finding extends findings from existing literature such as Duvvuri et al. (2007) and Hruschka (2013) which report that estimations are biased in small scale models. Findings of this study suggests that, not only is the estimation is possibly biased in different scales, but also the cross effect partners may be misidentified. A large-scale cross effect model provides a more complete view of the interdependency between many categories.

The model also provides information about whether a category is more self-price-determined or cross-price-determined. For example, the results show that the utility of frosting is more influenced by the price of its cross effect partner, rather

than its own price. The other three categories are self-determined in that their utilities are more influenced by their own prices. It implies that there are autonomous categories and complementary categories. Large-scale cross effect analysis is a promising method to identify such categories. Existence of relationships between seemingly unrelated entities requires an open mind. However, existing marketing literature shows promise. For example, Carpenter et al. (1994) find that product attributes that are not intuitively relevant to consumers' decision making can actually influence consumers' decision making.

6.3 Data Sparseness and Its Impact on the MVP and ANN Models

Data sparseness will invalidate any data-driven model. The MVP model will simply give an unpredictable outcome when there is not enough evidence for a relationship.

ANN requires good training samples to have a good learning outcomes. By good, it is meant that the model must be consistent with frequent reinforcement of the correct relationship. In this sense, ANN fits best to the relationship that is complicated, but not profound. But, with insufficient correct examples to show to NN, effective learning will not happen and the ANN is easily degraded by noisy samples. In other words, in a sparse data situation, labeling a noisy sample to differentiate it from interesting samples may be important to NN training. At this time, there is not a tactic that has been developed on this matter in business research. Data sparseness is very common in business research. Thus, this research contributes to machine learning applications in business.

6.4 Non-linear Relationship, and the General Effect of an ANN Input

The non-linear relationship provided by ANN has managerial implications, but is also a new approach for identifying regional effects of IVs. Regional effects indicate that the effect of an IV is not constant along the IV's value range, but rather the effect is relatively constant in a region of the IV's value range. Knowing regional effects provides better understanding of the IV's effect. Regression models assume a general effect of IVs on DVs, and do not directly provide regional effect estimations. Research about moderation and mediation is related to regional effect in that it studies how the level of an IV's effect is influenced by the value of another IV. However, if the IV itself has a regional effect, a regression model captures only its general effect. Creating dummy variables to represent regions require knowledges of the regions before estimation.

The MVP model is a regression model. Just as the way that a regression coefficient is interpreted, the estimated effect of an IV in the MVP model is assumed independent from that of other IVs. This assumption allows the expression of an IV's general value. It is valid to make a prediction of marginal change of DV by providing value change of a single IV. In contrast, ANN does not assume independence between IVs' effects. Thus, there is no specification of the general effect of an IV, but rather the effects of IVs are always interdependent. Thus, there is not prediction on a DV change by only providing one single IV value because ANN needs a combination of IVs to make a prediction of DV value.

The outcome is that the MVP has a parameter matrix; but the ANN does not. In essence, the ANN relaxes an assumption that IVs are independent from each other. This feature is very useful in situations where IVs are not independent.

6.5 Complexity of MVP and ANN models

Fitting data to both the MVP and ANN models requires a different mindset. In general, ANN requires relatively less effort. Ease of specifying ANN falls in three areas. Model construct is relatively easier to understand because it follows normal flow of learning, thoughts of training and testing, and the logic of human intelligence. When presented in graph form, the ANN model is easy to understand as a process of layered feeding forward data flow. Further, it does not rely on probability distribution theory. It uses a logistic function for value transformation purposes only. The user can initiate ANN just by remembering some rules of setting up parameter values. Finally, the implementation is easier in that programmers can take the algorithm description and write program code to run it. In contrast, using the MVP model requires deep understanding of probability distribution theories such as multivariate normal, Bayesian statistics, random sampling and the Markov Chain model.

Chapter 7. Conclusions, Limitations and Future Research

7.1 Conclusions and Limitations

This dissertation extends cross effect research in two areas.

The first is extension to the big data analytics context. Cross effect research studies interactions of latent utilities that influence consumer behavior. It has high relevancy in business. With the trend of big data analytics, this research can play an increasingly important role in terms of modeling consumer behavior by integrating large-scale categories. Most of the existing MVP models take the approach of pre-specifying cross effect partners. To advance model development, this study relaxes that constraint and allows evidence-based pairing of cross effect partners. This approach technically needs many more tests of possible cross effects and, thus, can become computationally cumbersome. But, the benefit is that it fits the paradigm of a data-driven approach. The data analysis, reported in previous sections, shows that pre-specified cross effect partners based on prior assumptions can be very different from the partners identified by data evidence.

Second, an alternative approach, ANN modeling, is examined. ANN's construct and learning mechanisms are customized to fit the specific problem. The ANN model fits the cross effect context because the nature of cross effect analysis is a three layer decision making model (recall Figure 3.3). Using far less time and resources, the ANN model is able to finish computations and have prediction performance similar to that of the MVP model. Additionally ANN can be used for large-scale cross effect analysis where MVP models cannot be used.

The advantages of ANN include faster computation, prediction orientation, easy to implement, and providing information on non-linear relationships between IVs and DVs. A disadvantage is that a general effect of each IV on a DV cannot be extracted (recall section 6.4 for discussion of this issue). This limitation goes against conventional expectations of the need for understanding an IV's general effect. Another disadvantage is that an ANN model may be less robust because it is case driven. The prediction performance of a trained ANN depends on quality and quantity of the training samples.

This study has several limitations as described below.

First, this study extends the MVP model. Its performance in large-scale cross effect is examined. This study finds that, compared with the ANN model, the MVP model has disadvantages in computation time and resource requirements. The main reason is that the MVP model relies on the MCMC method and needs to utilize the vectorization operations to improve computational performance.

A limitation is that the other econometric model, multivariate logistical model (MVL), used in the cross effect literature has not been compared in this dissertation. The MVL model usually does not assume multivariate normal distribution, but uses customized softmax probability such as the model in Russell and Petersen (2000). The Russell and Petersen (2000) model has a deterministic estimation process and, thus, avoids the requirement of calculating high dimension integrals. Compared with the MVP model, the MVL does not explicitly take into account interdependence of purchase utility among categories, but rather treats the cross

effect as an outcome of the conditional choice probability model. This dissertation takes the perspective that the cross effect phenomenon is more about interactions in latent purchase utilities than about dependency of purchase decisions. Section 2.4 provides detailed discussion. In short, the Russell and Petersen (2000) model is constrained to cross effects that occur only when the partner category is purchased and, thus, does not allow "informational" cross effect.

The main objective of this study is to introduce and examine large-scale cross effect. A direction of future research identified by this study is to extend the heterogeneity analysis in large-scale effects. A conceptual model extending heterogeneity research is advanced for this purpose and is described in section 7.2.2. The extracted data in this dissertation turn out to be not sufficient for a household-level heterogeneity analysis. Practically, the data sparseness problem largely constrains the applicability of household level analysis (recall section 3.1.3), and single level models with extended capability are more practically useful (Duvvuriet al. 2007). Considering the scope of this dissertation, the testing of the conceptual model is left for future research.

7.2 Future Research

There are several specific future research directions.

7.2.1 Theory of Cross Effect between Unfamiliar Pairs of Categories

A theoretical question that remains unanswered in cross effect research is whether cross effect is possible between seemingly unrelated categories. For example, Ainslie and Rossi (1998, p. 94) question the approach of modeling

unfamiliar categories, and express the skepticism that change in ketchup prices would impact the demand of canned tuna.

Some studies follow this line of thought. For example, (Manchanda et al. 1999) did not find significant cross effect between such a unfamiliar pair as cake mix and detergent. However, the way they did it is different from the large-scale model in this study. They first run a small scale pre-test to verify that there is no cross effect between cake mix and detergent. Then, in the following models, they constraint the cross effect as being none. Results of this study reported in this dissertation, as may be expected, show that behaviors of cross effect can be very different when estimated in a small vs. large-scale datasets.

Some studies have an open mind and search for alternative findings. For example, Duvvuri et al. (2007) find that consumers can be sensitive to price change in one category, but at the same time not sensitive to that of the complementary category. Their results are consistent among all the three pairs of categories, cake mix and frosting, detergent and softener, and spaghetti and sauce. Their findings can be an indicator of very strong budgeting effect, because consumers' response to price change leads to spending change, and that spending change must be compensated by adjusting spending on another category under a constraint budget. In such a case, a budget-sensitive consumer facing a price increase of ketchup up can result in a decreased purchase of tuna fish. The case explained above is theoretically possible and an expected observation because of the strength of budgeting effect on consumption behavior.

7.2.2 The Spending Habit Heterogeneity Model and Propositions

Combining theoretical discussions in section 2.5 and 3.1, this study forms a conceptual model of spending habit in cross category decision making.

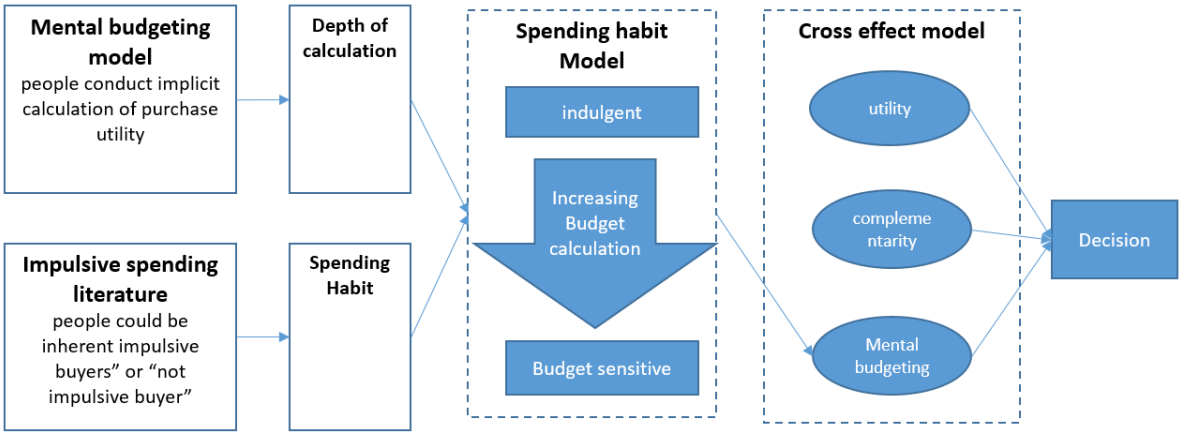


Figure 7.1 Conceptual model of spending habit heterogeneity

This model extends the mental budgeting component in the original cross effect model. The mental budgeting component is examined by Duvvuri et al. (2007). This study extends the component and introduces the concept of depth of budget calculation and spending habits. This model theorizes the presence of mental budgeting component of consumers' decision making processes in a cross effect context. The existing literature focuses on complementarity and utility, but the mental budgeting component has been ignored.

This model generates several propositions. In section 2.5.3, a variable is proposed to measure consumer's spending habit level (i.e., variance of total spending per trip). The higher the variance the lower the mental budgeting calculation. The two propositions about impact of mental budgeting on cross effect are:

Proposition-1: variance of trip total spending is negatively related to consumers' cross price effect.

Proposition-2: variance of trip total spending is positively related to consumers' cross promotion effect.

Lower mental budgeting is reflected by higher variance of spending. When mental budgeting is low, the consumers are less sensitive to budget overdraw. Suppose categories A and B are cross effect partners and the cross effect is negative and significant. At a price drop of A, the consumers are likely to buy more A. The cross effect predicts that they buy more B as well. It may result in consumers overspending. In this chain of logic, then whether the consumers who buy more B will also depend on their mental budgeting level. If the consumers are low mental budgeting ones, they are more likely to buy B. In contrast, when mental budgeting is high, the consumers tend to buy only A, the discounted category, because buying the other category will overdraw the budget. Similar logic is applied to cross effect of promotion.

7.2.3 ANN Incorporating Existing Knowledge

ANN learns everything from the training data. This fact, on one hand, utilizes data-driven discovery, but on the other hand, allows biased outcomes if the data are seriously contaminated. It also burdens the learning procedure because the existing knowledge has to be re-learned from data. One possible research direction is to customize ANN for the purpose of enabling knowledge embedding.

For example, the hierarchical relationship between IVs can be pre-specified before ANN training.

Another direction is to focus on business knowledge discovery. ANN application in hard science mainly is to close the gap between artificial intelligence and human intelligence, such as recognizing object concept from digital image, and natural language processing. In business research, it is more interesting to discover unseen patterns. ANN's application is thus, to extend human intelligence capability. In short, if ANN in hard science is to make human-like machines, then ANN in business is to make super intelligent humans who can read reports 1000 times faster, and engage learning patterns in seconds by reading through big data rather than in years by life experience.

7.2.4 Evolutionary Learning Algorithm

Gradient descent learning is used to train the ANN model. An alternative learning method is the evolutionary algorithm. Baczyński and Parol (2004) point out that the gradient descent algorithm is subject to the trap of local solutions, especially for training multi-layer ANN models. The large-scale cross effect model contains multiple outputs that are expected to be correlated and it contains multiple sources of inputs that are expected to interact. Generally, the context of large-scale cross effect is complex and, thus, the surface of the error function can be bumpy because the error function can reflect the complexity of the learning problem. In business research, ANN techniques have not been widely studied in the cross effect context. In this study, a basic ANN model is demonstrated with the commonly used gradient descent learning algorithm. A future research opportunity

is to train ANN with evolutionary learning algorithms and examine the impact on prediction performance.

7.2.5 Impact of Data Preparation on Research Findings and Drawing Conclusions

Academic business research has been focused on theoretical advancement; and the impacts of data preparation on research findings have not been attractive to research efforts. However, the trend of big data analytics calls for research in this area. When preparing a dataset there are places where choices have to be made for excluding certain types of data. The impact of such decisions on the resulting dataset, and the impact of this dataset on analysis outcomes and on conclusions drawn, have not been studied fully- except in the investigation of the category scale effect. Several research questions can be asked in this area. For example, consumers entered in a dataset have to meet the requirement that they must make at least ten purchases of each category in the list. Choosing consumers with ten or more purchases and those with five or more purchases can generate two very different datasets. Then does the analysis lead to a different outcome? Does it lead to different conclusions to be drawn? Answering such questions can help managers avoid misleading analysis. It also help researchers avoid drawing misleading research conclusions.

7.3 General Conclusion

In general, this dissertation extends cross effect research and examines the MVP model and the ANN model in the large-scale cross effect context. The perspective of evidence-based knowledge discovery makes this dissertation fit into

the big data analytics research. The findings shed light on the model performance, operational feasibility, and prediction accuracy with increasing category scales. It demonstrates several techniques to customize the ANN model according to the feature of large-scale cross effect analysis. This study can spawn many future research directions as discussed in section 7.2.

Specifically, there are several major findings. First, in the large-scale cross effect context, the ANN model is more scalable than the MVP model. The MVP model can be estimated only in the scale 4 and 8, but the ANN can be computed in all the 4 scales, 4, 8, 16, and 32. Second, to properly measure prediction accuracy for different model scales, this study introduces the measures of hit rate lift and base categories hit rate. Both are normalized measures that can be used to compare models of different scales. Third, this study customizes ANN's configurations to make it fit to the large-scale cross effect analysis context such as the biased cross entropy error function. Fourth, this study finds that, in general, the base 4 categories prediction hit rate is better in larger scale models such as in scale 8 (both MVP and ANN) and 16 (ANN only) models. But when the model scale is too large, such as 32 (ANN only), the estimated prediction model becomes useless in terms of prediction because it cannot reach the hit rate from a naïve prediction. Even though using 32 hidden nodes can largely increase the ANN model's capability, the resulting model performance still cannot outperform smaller scale models.

Generally, the large-scale cross effect analysis fits the big data analytics paradigm that emphasizes evidence-based problem solving. Applications of the

MVP and the ANN model have their own advantages and disadvantages. With large and rich datasets becoming increasingly available, research on and applications of large-scale models and techniques are highly relevant from both academic and industrial perspectives.

Appendix

Appendix A.1 Representative cross effect (CE) literature

Literature	Categories	Heterogeneity	General model
(Manchanda et al. 1999)	Grocery (Cake mix, frosting), detergent, softener in chain stores	Random effect Captured by household demographic variables	$\mu_{hJt} = \beta_{hJ0} + \beta_{hJ1} * Own\ Effect + \beta_{hJ2} * Cross\ Effect + \varepsilon_{hJt}$ $\varepsilon_{ht} \sim MVN(\mathbf{0}, \Sigma)$ CE: $\beta_h = D_h * \mu + \lambda_h, h = 1\ to\ H$ $\beta_h = \{\beta_{h0}, \beta_{h1}, \beta_{h2}\}, \lambda_h \sim MVN(\mathbf{0}, \Lambda)$
(Russell et al. 2000)	Grocery 4 categories of paper products	A fixed effect model of CE heterogeneity because CE is estimated each HH k is estimated*	$U_{ikt} = \beta_i + HH_{ikt} + MIX_{ikt} + \sum_{j \neq i} \theta_{ijk} * C_{jkt} + \varepsilon_{ikt}$ $C_{jkt} \rightarrow purchase\ or\ not\ purchase\ of\ j$ $MIX_{ikt} = \gamma_i * \log(PRICE_{ikt})$ $HH_{ikt} = \delta_1 * \log(TIME_{ikt} + 1) + \delta_2 * LOYAL_{ik}$ ε_{ikt} ~ extreme value distribution. CE: $\theta_{ijk} = \delta_{ij} + \phi * SIZE_k$
(Chib et al. 2002)	12 grocery items in a "typical" basket	Fixed effect Captured by a household specific constant term (b_h) and a household/category specific constant term (c_{hj})	$Z_{htj} = X'_{htj}\beta_j + b_h + c_{hj} + \varepsilon_{htj}$ heterogeneity: b_h, c_{hj}
(Li et al. 2005)	Choices of Financial investment products	Random effect regressing parameters are regressed on household demographic and social status variables	$U_{ijt} = \beta_i \lvert O_j - DM_{jt-1} \rvert + \gamma_{1ij} COMPET_i + \gamma_{2ij} OVERSAT_i + + \gamma_{3ij} SWIT_{it} + \varepsilon_{ijt}$ Heterogeneity: $\beta_i = \mu_0 + \mu_1 * EDUCAT_i + \mu_2 * SEX_i + \mu_3 * AGE_i + \mu_4 * INCOME_i + e_i$ $\gamma_{ki} = \omega_{0k} + \omega_{1k} * EDUCAT_i + \omega_{2k} * SEX_i + \omega_{3k} * AGE_i + \omega_{4k} * INCOME_i + \xi_{ki}$
(Wedel and Zhang 2004)	3 subcategories of orange juice	--	$\ln(q_{r,t,c}) = \mu_{r,c} + \ln(p_{r,t}) * A_{r,c} + x_{r,t,c} * \Gamma_{r,c} + s_{r,t} * k_c + \psi_c(t) + \varepsilon_{r,t,c}$

(Song and Chintagunta 2006)	2 categories of detergent and 2 categories of softener in 50 chain stores	Address the inter-collinearity with instrumental whole sale prices	Derived from (Chib et al. 2002)
(Duvvuri et al. 2007)	ACNielsen, 6 categories grocery, HH made at least one purchase in each of the 6 categories	Same as (2), but with different variables	Utility for J categories: $\boldsymbol{u_{it} = \alpha_i + X_{it} * \beta_i + \epsilon_{it}}$ $\boldsymbol{X}$ ={price, promotion, inventory} $\boldsymbol{\epsilon_{it} \sim MVN(0, \Sigma_u)}$ $\boldsymbol{\alpha_{it} \sim MVN(Z_i \alpha, \Sigma_a)}$ $\boldsymbol{\beta_i \sim MVN(\mu_\beta, \Sigma_\beta)}$
(Boztuğ and Reutterer 2008)	Large number of grocery categories	--	$\boldsymbol{U_{int} = \beta_i + \delta_{1i} * \ln(TIME_{int} + 1) + \delta_{2i} * \mathrm{LOYAL}_{in} + \gamma_i * \ln(PRICE_{int}) + \xi_i * DISPLAY_{int} + \sum_{j \neq i} \theta_{int} * C_{jnt} + \varepsilon_{int}}$

Appendix A.2 Pair-wise Joint purchase frequency

* Cells value larger than 99 are highlighted
* Read from row perspective:
  row y1 in column y1 is 2451, the total number of trips that category 1 is purchased
  row y1 in column y2 is 963, the total number of trips that both category 1 and 2 are purchased
  Thus, 963 is part of 2451.

	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11	y12
y1	2451	963	176	56	69	40	30	74	258	121	67	18
y2		1685	118	45	52	28	20	52	228	86	48	14
y3			3147	337	90	61	16	76	118	115	49	9
y4				1109	40	20	17	28	54	35	22	5
y5					1049	35	9	61	46	50	25	8
y6						855	6	28	49	25	22	3
y7							288	21	20	14	12	8
y8								1008	43	50	27	7
y9									1855	89	78	29
y10										1340	86	16
y11											692	11
y12												217

	y13	y14	y15	y16	y17	y18	y19	y20	y21	y22	y23	y24
y1	51	122	93	63	31	7	33	123	7	17	12	94
y2	33	137	94	45	17	8	21	71	2	12	14	64
y3	31	64	17	80	19	7	20	50	9	13	29	274
y4	14	25	8	38	4	4	11	21	2	15	34	124
y5	22	21	9	39	11	1	9	26	5	7	10	44
y6	11	22	4	33	4	1	3	14	1	5	1	42
y7	7	15	2	14	5	3	8	20	3	0	0	22
y8	7	39	12	21	15	6	20	33	4	7	8	37
y9	37	140	19	60	13	2	18	72	5	12	17	90
y10	26	53	17	63	8	5	8	35	4	11	13	74
y11	22	63	5	30	5	2	4	20	3	7	8	34
y12	4	14	1	9	3	0	3	7	0	0	3	8

	y25	y26	y27	y28	y29	y30	y31	y32
y1	28	59	4	16	38	41	10	5
y2	22	41	4	17	23	32	7	1
y3	82	198	17	98	137	111	45	23
y4	36	76	9	64	73	55	19	8
y5	10	34	3	7	11	15	5	1
y6	8	27	7	3	7	12	4	3
y7	3	12	0	0	0	6	2	0
y8	7	26	2	7	18	8	4	3
y9	26	54	6	13	24	26	10	5
y10	17	57	6	20	20	14	9	2
y11	12	31	2	7	14	16	5	1
y12	2	3	1	1	1	2	1	1

	y13	y14	y15	y16	y17	y18	y19	y20	y21	y22	y23	y24
y13	478	19	5	11	9	2	6	19	3	6	2	24
y14		880	17	44	7	3	14	42	2	7	13	43
y15			378	10	5	4	12	18	4	5	2	12
y16				934	5	2	8	24	5	9	11	53
y17					307	0	4	16	9	3	1	11
y18						79	8	19	1	0	0	4
y19							326	39	5	8	5	23
y20								1009	10	15	2	36
y21									138	20	0	5
y22										305	5	19
y23											362	41
y24												1759

	y25	y26	y27	y28	y29	y30	y31	y32
y13	6	20	2	8	3	5	3	1
y14	13	23	0	10	5	12	4	2
y15	3	11	2	2	2	4	1	1
y16	11	40	2	5	17	12	8	6
y17	2	7	1	3	3	3	0	0
y18	2	3	1	0	0	0	0	0
y19	0	7	0	0	5	2	2	1
y20	6	24	2	8	15	8	8	3
y21	2	2	1	0	2	4	3	0
y22	4	9	3	4	3	3	5	1
y23	13	27	2	10	23	17	0	0
y24	43	139	22	45	65	48	28	3
y25	501	34	3	15	23	25	11	11
y26		1264	44	40	44	39	24	4
y27			152	5	2	2	5	2
y28				402	16	26	8	5
y29					491	18	7	2
y30						627	10	1
y31							257	4
y32								168

Appendix A.3 Price parameters draws of the four base category

The four plots shows the price parameter draws for the four base categories.

The plot at top-left is for cake mix, y1,
top-right is for frosting, y2,
bottom-left is for detergent, y3, and
bottom-right is for softener, y4.

Each plots contains four sets of draws as shown by the legend. The series of "1" represents the effect of category-1 price on purchase utility.

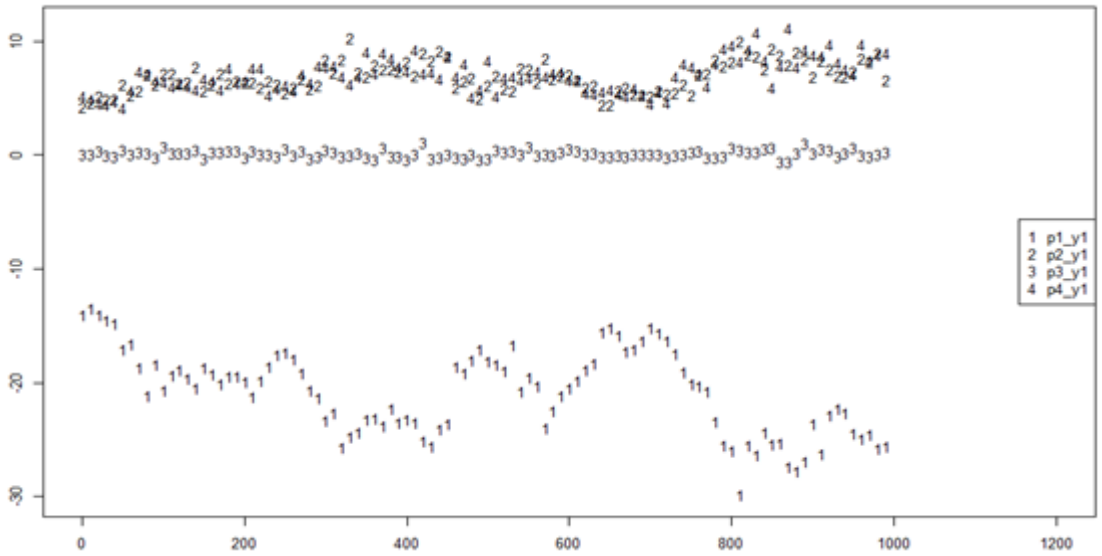
For the purpose of readability, the plots show only 100 random samples of total 1000 draws (took every 20^{th} of total 20000 draws).

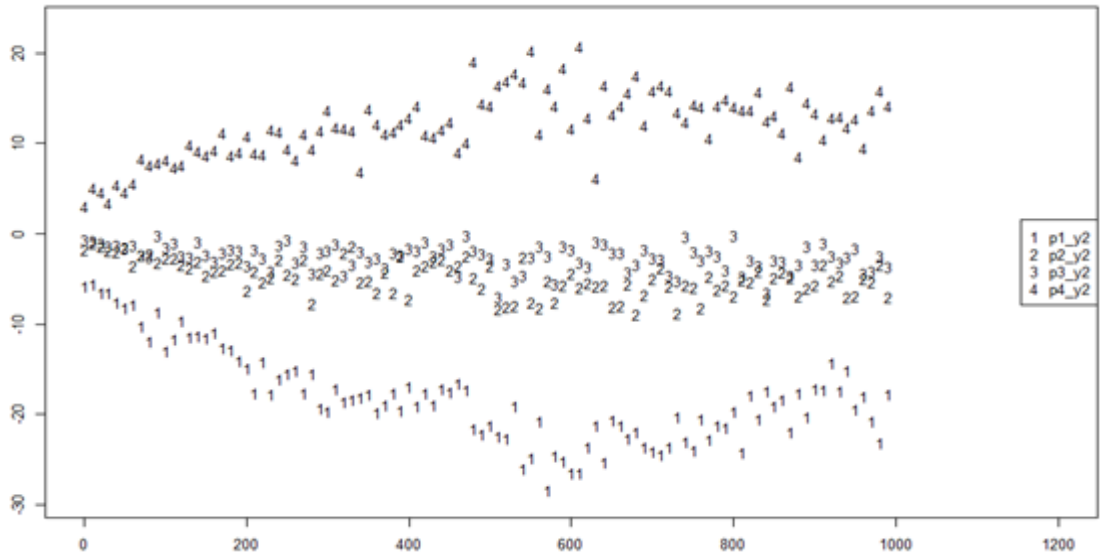
Y axis is the value of parameter

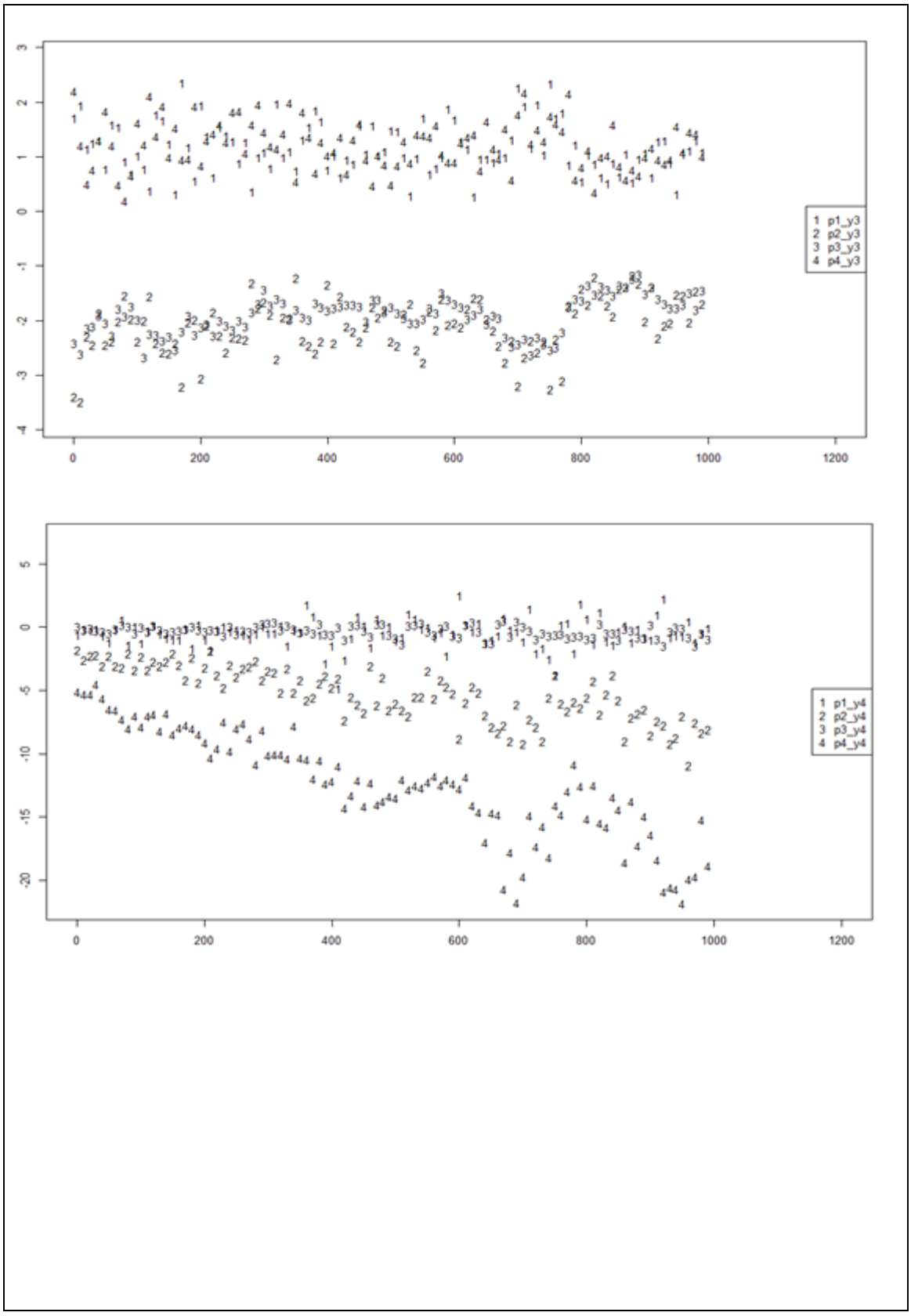
X axis is indices of draws.

R is number of rounds to run

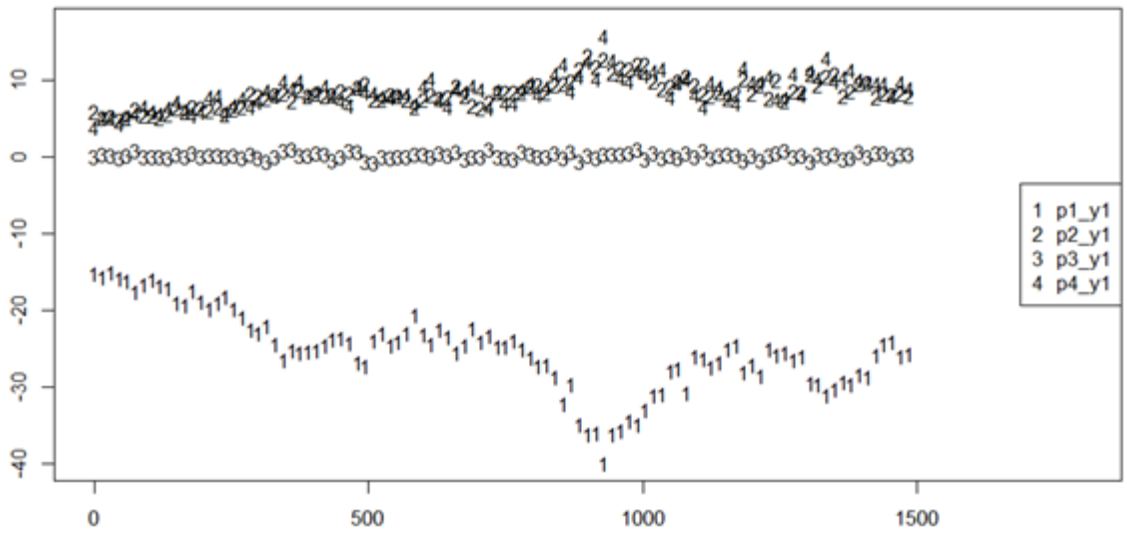
R=20000

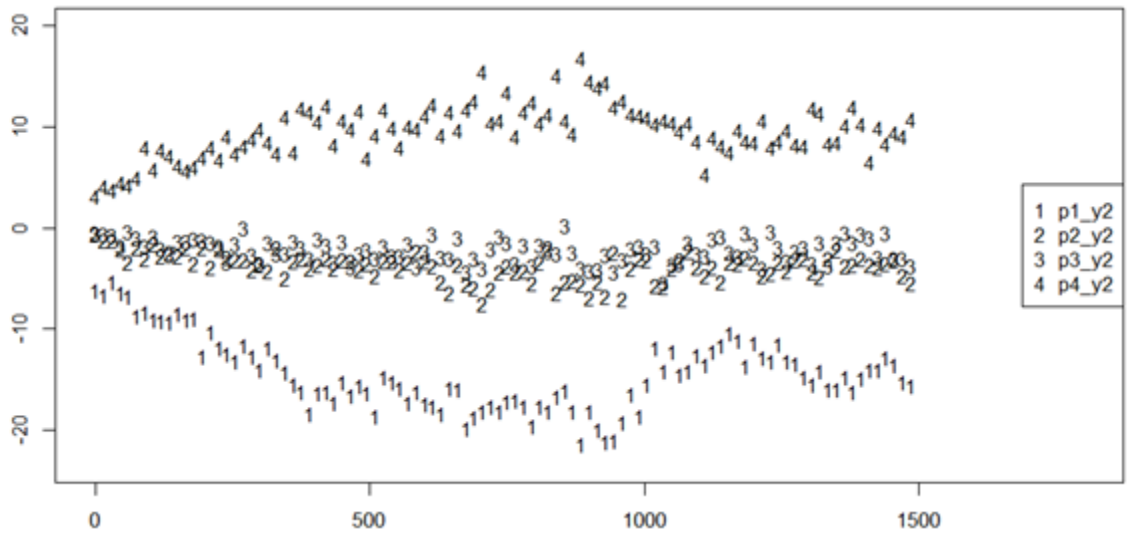

10
0
-10
-20
-30
0
200
400
600
800
1000
1200
1 p1_y1
2 p2_y1
3 p3_y1
4 p4_y1

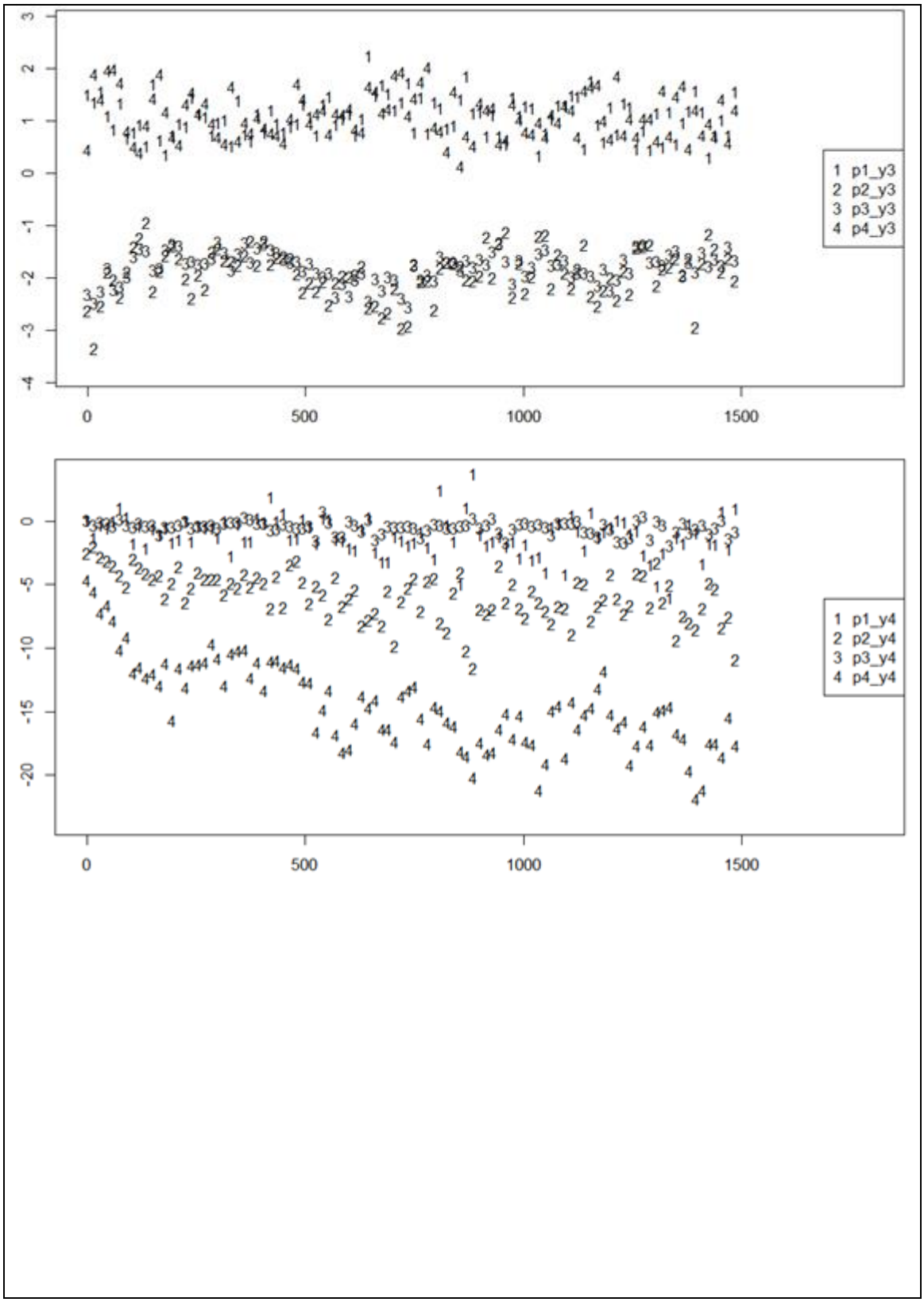

20
10
0
-10
-20
-30
0
200
400
600
800
1000
1200
1 p1_y2
2 p2_y2
3 p3_y2
4 p4_y2


1 p1_y3
2 p2_y3
3 p3_y3
4 p4_y3
1 p1_y4
2 p2_y4
3 p3_y4
4 p4_y4

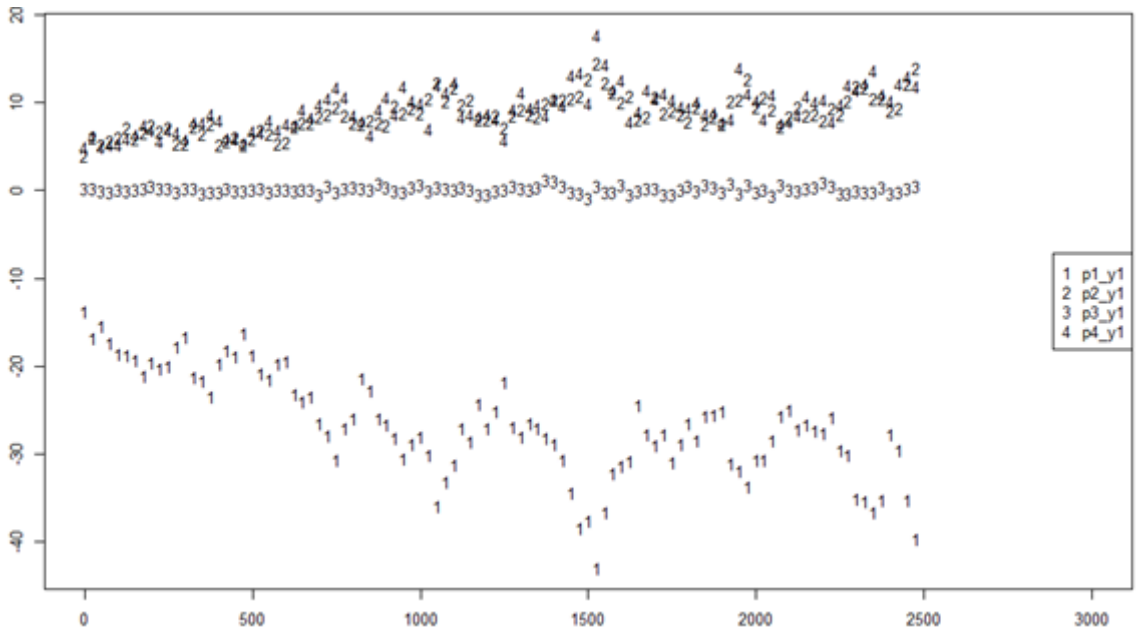
R=30000

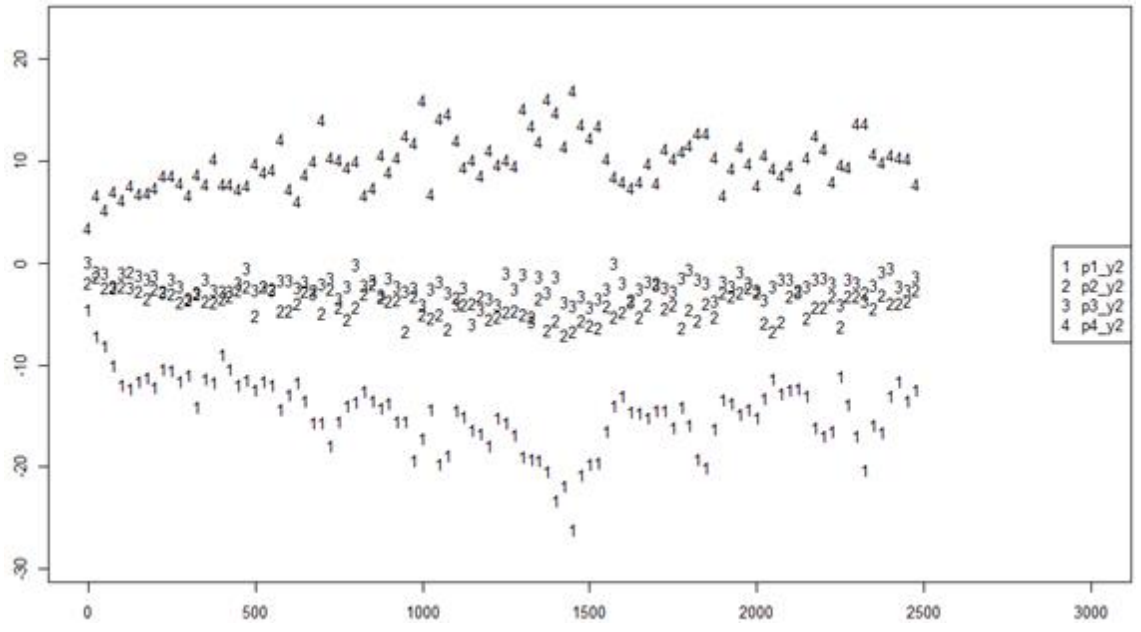

10
0
-10
-20
-30
-40
0
500
1000
1500
1 p1_y1
2 p2_y1
3 p3_y1
4 p4_y1

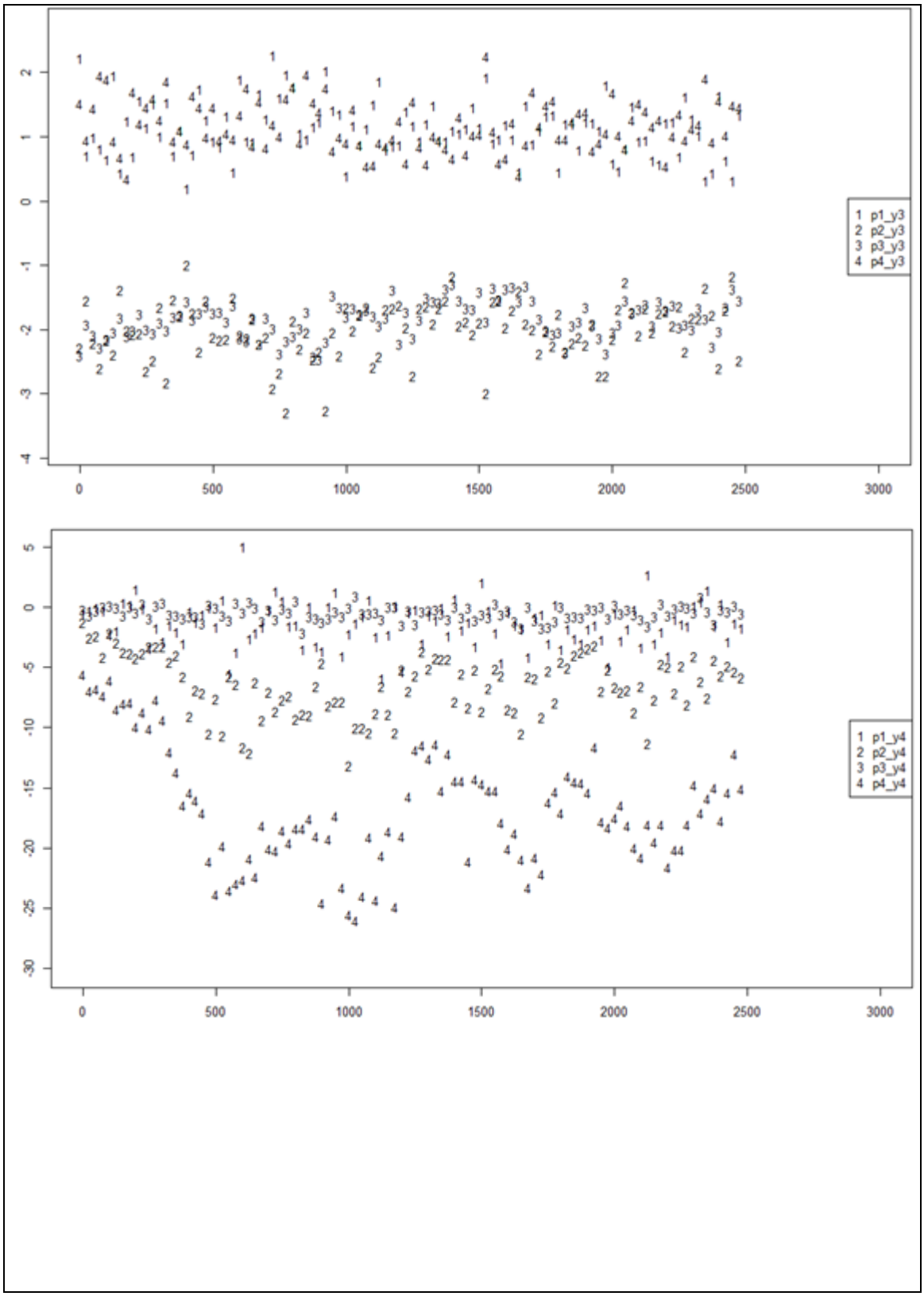

20
10
0
-10
-20
0
500
1000
1500
1 p1_y2
2 p2_y2
3 p3_y2
4 p4_y2


1 p1_y3
2 p2_y3
3 p3_y3
4 p4_y3
1 p1_y4
2 p2_y4
3 p3_y4
4 p4_y4

R=50000


20
10
0
-10
-20
-30
-40
0
500
1000
1500
2000
2500
3000
1 p1_y1
2 p2_y1
3 p3_y1
4 p4_y1


20
10
0
-10
-20
-30
0
500
1000
1500
2000
2500
3000
1 p1_y2
2 p2_y2
3 p3_y2
4 p4_y2


1 p1_y3
2 p2_y3
3 p3_y3
4 p4_y3
1 p1_y4
2 p2_y4
3 p3_y4
4 p4_y4

Appendix B Inputs dependency in ANN

The equations below provides a simple form of an ANN with 2 input nodes, x_a, x_b, 2 hidden node h_1, h_2, and one output node y, with logit activation function.

$$h1 = \text{logit}\ (w^{(x1h1)} * x_a + w^{(x2h1)} * x_b)$$

$$h2 = \text{logit}\ (w^{(x1h2)} * x_a + w^{(x2h2)} * x_b)$$

$$Y = \text{logit}\ [\ w^{(h1y1)} * h1 + w^{(h2y1)} * h2]$$

Assume that x_a is household information IV, x_b is price IV, and y is DV of probability to purchase. The main effect is x_b on y. The heterogeneity is captured by moderation effect of x_a on the main effect. In the ANN model shown above, the effect of x_b on y is the integrated path value of $w^{(x2h1)}$, $w^{(x2h2)}$, $w^{(h1y1)}$ and $w^{(h2y1)}$. Even though the path value of $w^{(x2h1)}$ and $w^{(x2h2)}$ is independent from the value of x_a, the path value of h_1 and h_2 are dependent on value of x_a. In such a case, the effect of x_b on y is not independent from the value of x_a. Putting the 3 equations together forms:

$$\Pr(y = 1|X) =$$

$$LOGIT\ [\frac{w^{(h1y1)}}{1 + e^{-[\ w_0 + w^{(x1h1)} * x_a + w^{(x2h1)} * x_b\ ]}} + \frac{w^{(h2y1)}}{1 + e^{-[\ w_0 + w^{(x1h2)} * x_a + w^{(x2h2)} * x_b\ ]}}].$$

This model is capable of learning a relationship that the effect of x_b to y is dependent on the value of x_a. For example, the impact of x_b on $\text{Prob}(y = 1)$ can

be high when right hand side LOGIT function is around the value 0.5; but the effect would be low when the LOGIT is around 0 or 1. The LOGIT value is dependent on the value of x_a.

Appendix C Mean of Percentage Error (MPE)

A measure of model prediction performance is defined as shown in equation (6.1), i.e., the Mean of Percentage Error (MPE).

$$\frac{\mathbf{1}}{\boldsymbol{B}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{B}}\frac{|\boldsymbol{t_b}-\boldsymbol{p_b}|}{\boldsymbol{p_b}}\ , \qquad \boldsymbol{b}\ \in \textbf{B types of basket in a prediction profile} \qquad (\mathbf{6.1})$$

Given a dataset of B baskets that t_b is the true number of purchases of a basket $b \in B$. The term p_b is the prediction of t_b from a prediction model.

The MPE is a slightly revised version of the Mean Absolute Percentage Error (MAPE) described in (Armstrong and Collopy 1992). The MAPE can be written as

$$\frac{1}{B}\sum_{b=1}^{B}\frac{|t_b-p_b|}{t_b}, \qquad MAPE\ of\ (\text{Armstrong and Collopy 1992})$$

According to (Armstrong and Collopy 1992), the features of MAPE are (1) unit-free which cancel out the effect of large unit over small unit of t_b, (2) heavier penalty on case of $p_b > t_b$ than that of $p_b < t_b$. Our MAE measure replaces the denominator of MAPE from t_b to p_b. This adjustment makes MPE punishing false negative predictions much more than false positive predictions. This feature fits marketing context better because, in general, false negative of prediction is more costive than false positive. For example, false negative can lead to losing a sales opportunity because of failure to identify a customer; while false positive may lead to sending mails to unresponsive customers because of mistakenly identifying a customer. At many times, the former case is more costive to business.

Here is an example to illustrate how MPE works. Suppose there is only one basket. If the basket has true purchases of 1000, and the prediction is 500, then the MPE is |1000 – 500|/500 = 1. In contrast, if purchases is 500, but prediction is 1000, then the calculation has |500 – 1000|/1000 = 0.5. The former has a higher MPE because missing 500 sales opportunities is heavier punished than over-predicting by 500 sales. Beside this feature, this measure takes off scale effect (unit-free in (Armstrong and Collopy 1992)).

When a prediction is zero, the MPE will run into dividing by zero error. To avoid it, cases of prediction zero are omitted from MPE calculation. First of all, by examining the prediction outcomes, we find that prediction of zero happens at most times on cases of true zero. The cases that prediction is zero and true number is not zero are very rare, and when it happens, the true number at most times are 1. When the true number is zero, ignoring these cases makes no differences to MPE because the prediction error is zero anyway and should be ignored from MPE. When true number is not zero, ignoring these cases will less count prediction errors. However because number of cases is very small, and the prediction error made is also small, ignoring them will not largely change the result of model comparison using MPE.

Taking average over number of basket types, $\frac{1}{B}$, makes it category scale free that models' performance of four categories, 8 categories and 16, 32 categories can be compared. This measure is used to compare general model

prediction performance in the increasing category scales in both the study of MVP model and ANN model.

After ignored the predictions of zero, the range of the MPE measure is zero to positive infinity. Zero means a perfect prediction that made no mistakes. Infinity error happens when the true number is very large but the prediction is very small. For example, a basket is purchased 100 thousand times in transaction databases, but is predicted 1 purchase only. Then error is 99999. Such a large error simply means that the model makes very inaccurate prediction. We can roughly interpret it as number of predictions mistakenly made by the model for each prediction of purchase of a basket. With such an interpretation, a reasonable range of MPE is from 0, prefect prediction, to 1, very bad prediction.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Aguinis, H., Forcum, L. E., & Joo, H. (2013). Using Market Basket Analysis in Management Research. *Journal of Management* 39 (7): 1799–1824.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1), 57-78.

Allenby, G. M., Rossi, P. E., & McCulloch, R. E. (2005). "Hierarchical Bayes Models: A Practitioners Guide." *SSRN working paper,* Ohio State University, University of Chicago.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1), 69-80.

Baczyński, D., & Parol, M. (2004). Influence of artificial neural network structure on quality of short-term electric energy consumption forecast. *IEE Proceedings-Generation, Transmission and Distribution*, *151*(2), 241-245.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*(1), 191-211.

Baumeister, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of Consumer Research*, *28*(4), 670-676.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41-48). ACM.

Bergstra, J., Bastien, F., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., ... & Bengio, Y. (2011). Theano: Deep learning on gpus with python. In NIPS 2011, *BigLearning Workshop, Granada, Spain*.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press

Boxall, P. C., & Adamowicz, W. L. (2002). Understanding heterogeneous preferences in random utility models: a latent class approach. *Environmental and resource economics*, *23*(4), 421-446.

Boztuğ, Y., & Hildebrandt, L. (2003). *A Market Basket Analysis Based on the Multivariate MNL Model* (No. 2003, 21). Discussion papers of interdisciplinary research project 373.

Boztug, Y., & Reutterer, T. (2008). A combined approach for segment-specific market basket analysis. *European journal of operational research*, *187*(1), 294-312.

Carpenter, G. S., Glazer, R., & Nakamoto, K. (1994). Meaningful brands from meaningless differentiation: The dependence on irrelevant attributes. *Journal of Marketing Research*, 339-350.

Ceperley, D., Chen, Y., Craiu, R. V., Meng, X. L., Mira, A., & Rosenthal, J. (2012). Challenges and advances in high dimensional and high complexity monte carlo computation and theory. *Banff International Research Station*.

Chib, S., Seetharaman, P. B., & Strijnev, A. (2002). Analysis of multi-category purchase incidence decisions using IRI market basket data. *Advances in Econometrics*, *16*, 57-92.

Chung, J., & Rao, V. R. (2003). A general choice model for bundles with multiple-category products: Application to market segmentation and optimal pricing for bundles. *Journal of Marketing Research*, *40*(2), 115-130.

Cooper, J. C. (1999). Artificial neural networks versus multivariate statistics: an application from economics. *Journal of Applied Statistics*, 26(8), 909-921.

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*, *24*(4), 595-615.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science*, *52*(4), 597-612.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... & Ng, A. Y. (2012). Large scale distributed deep networks. In *Advances in Neural Information Processing Systems* (pp. 1223-1231).

Dippold, K., & Hruschka, H. (2013). A parsimonious multivariate poisson model for market basket analysis. *Review of Managerial Science*, *7*(4), 393-415.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5), 352-359.

Duvvuri, S. D., Ansari, A., & Gupta, S. (2007). Consumers' price sensitivities across complementary categories. *Management Science*, *53*(12), 1933-1945.

Fiebig, D. G., Keane, M. P., Louviere, J., & Wasi, N. (2010). The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science*, *29*(3), 393-421.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research 5*(2), 103-123.

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R Journal*, *2*(1), 30-38.

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, *15*(1), 55-86.

Heath, C., & Soll, J. B. (1996). Mental budgeting and consumer decisions. *Journal of Consumer Research 23*(1), 40-52.

Holt, D. B. (1997). Poststructuralist lifestyle analysis: Conceptualizing the social patterning of consumption in postmodernity. *Journal of Consumer research 23*(4), 326-350.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems*, 64, 130-141.

Hruschka, H. (1993). Determining market response functions by neural network modeling: A comparison to econometric techniques. *European Journal of Operational Research*, *66*(1), 27-35.

Hruschka, H. (2013). Comparing Small-and Large-Scale Models of Multicategory Buying Behavior. *Journal of Forecasting*, *32*(5), 423-434.

Hruschka, H. (2014). Analyzing market baskets by restricted Boltzmann machines. *OR spectrum*, *36*(1), 209-228.

Currim, I. S. (1981). Using segmentation approaches for better prediction and understanding from consumer mode choice models. *Journal of Marketing Research*, 301-309.

Intrator, O., & Intrator, N. (2001). Interpreting neural-network results: a simulation study. *Computational statistics & data analysis*, *37*(3), 373-393.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1), 39-59.

Kamakura, W. A., & Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, *26*(4), 379-390.

Kim, J., Allenby, G. M., & Rossi, P. E. (2002). Modeling consumer demand for variety. *Marketing Science*, *21*(3), 229-250.

Kim, Y., Street, W. N., Russell, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, *51*(2), 264-276.

Kline, D. M., & Berardi, V. L. (2005). Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, *14*(4), 310-318.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Kumar, U. A. (2005). Comparison of neural networks and regression analysis: A new insight. *Expert Systems with Applications*, *29*(2), 424-430.

Lendaris, G. G., Zwick, M., & Mathia, K. (1993). On matching ANN structure to problem domain structure. In *Proceedings of World Congress on Neural Networks' 93*.

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research*, *42*(2), 233-239.

Li, Y., & Ansari, A. (2013). A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science*, *60*(5), 1161-1179.

Manchanda, P., Ansari, A., & Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science*, *18*(2), 95-114.

McFadden, D. (1973): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontier of Econometrics*, ed. by P. Zarembka. New York: Academic Press.

McFadden, D. (1980). "Econometric models of probabilistic choice" in *Structural Analysis of Discrete Data*, ed. by C. Manski and D. McFadden. Cambridge, Mass.: M.I.T. Press.

McFadden, D. (1986). The choice theory approach to market research. *Marketing science*, 5(4), 275-297.

Mehta, N., & Ma, Y. (2012). A Multicategory Model of Consumers' Purchase Incidence, Quantity, and Brand Choice Decisions: Methodological Issues and Implications on Promotional Decisions. *Journal of Marketing Research*, *49*(4), 435-451.

Nielsen M. A. (2015), *Neural Networks and Deep Learning*. Determination Press.

Mishra, V. K., Natarajan, K., Padmanabhan, D., Teo, C. P., & Li, X. (2014). On theoretical and empirical aspects of marginal distribution choice models. *Management Science*, *60*(6), 1511-1531.

Moon, S., & Russell, G. J. (2008). Predicting product purchase from inferred customer similarity: An autologistic model approach. *Management Science*, *54*(1), 71-82.

Ng A., Ngiam J., Foo Y. C., Mai Y. & Suen C. (2010). UFLDL Tutorial. *http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial*

Niraj, R., Padmanabhan, V., & Seetharaman, P. B. (2008). Research Note-A Cross-Category Model of Households' Incidence and Quantity Decisions. *Marketing Science*, *27*(2), 225-235.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, *36*(1), 2-17.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of marketing research*, 134-148.

R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces*, *16*(3), 265-278.

Rojas, R. (2013). *Neural Networks: A Systematic Introduction*. Springer Science & Business Media.

Rook, D. W., & Fisher, R. J. (1995). Normative influences on impulsive buying behavior. *Journal of consumer research 22*(3), 305-313.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. John Wiley & Sons.

Rossi, Peter E. (2012). "*bayesm: Bayesian Inference for Marketing/Micro-econometrics.*" R package version 2.2-5.

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, *76*(3), 367-392.

Russell, G. J., Ratneshwar, S., Shocker, A. D., Bell, D., Bodapati, A., Degeratu, A., ... & Shankar, V. H. (1999). Multiple-category decision-making: Review and synthesis. *Marketing Letters*, *10*(3), 319-332.

Seetharaman, P. B., Chib, S., Ainslie, A., Boatwright, P., Chan, T., Gupta, S., ... & Strijnev, A. (2005). Models of multi-category choice behavior. *Marketing Letters*, *16*(3-4), 239-254.

Shechtman, O., Anton, S. D., Kanasky Jr, W. F., & Robinson, M. E. (2005). The use of the coefficient of variation in detecting sincerity of effort: a meta-analysis. *Work (Reading, Mass.)*, *26*(4), 335-341.

Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *The Journal of Marketing 21*(1), 3-8.

Song, I., & Chintagunta, P. K. (2006). Measuring cross-category price effects with aggregate store data. *Management Science*, *52*(10), 1594-1609.

Straughan, R. D., & Roberts, J. A. (1999). Environmental segmentation alternatives: a look at green consumer behavior in the new millennium. *Journal of consumer marketing*, *16*(6), 558-575.

Taylor-West, P., Fulford, H., Reed, G., Story, V., & Saker, J. (2008). Familiarity, expertise and involvement: key consumer segmentation factors. *Journal of consumer marketing*, *25*(6), 361-368.

Trochim, William M. *The Research Methods Knowledge Base, 2nd Edition*. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006).

Vohs, K. D., & Faber, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *Journal of consumer research*, *33*(4), 537-547.

Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, *43*(3), 303-343.

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The american statistician*, *50*(4), 284-293.

Wedel, M., & Kamakura, W. A. (1998). *Market segmentation: Conceptual and methodological foundations* (ed. 2). Kluwer Academic.

Wedel, M., & Zhang, J. (2004). Analyzing brand competition across subcategories. *Journal of Marketing Research*, *41*(4), 448-456.

West, P. M., Brockett, P. L., & Golden, L. L. (1997). A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*, *16*(4), 370-391.

Wind, Y. (1978). Issues and advances in segmentation research. *Journal of marketing research 15* (3), 317-337.

Wong, B. K., Bodnovich, T. A., & Selvi, Y. (1997). Neural network applications in business: A review and analysis of the literature (1988–1995). *Decision Support Systems*, *19*(4), 301-320.

Xu, M., Wong, T. C., & Chin, K. S. (2013). Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network. *Decision Support Systems*, *54*(3), 1488-1498.

Yang, S., & Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research*, *40*(3), 282-294.

Youn, S., & Faber, R. J. (2000). Impulse buying: its relation to personality traits and cues. *Advances in consumer research*, 27, 179-185.

VITA

Birth Place

Liaoning, China

Education

Master of Business Administration 2009
Missouri State University, Springfield MO, USA

Bachelor of Computer Application 2002
Liaoning Normal University, Dalian Liaoning, China

Professional positions

Teaching/Research Assistant 2010 to 2014
Gatton College of Business and Economics, University of Kentucky, Lexington KY, USA

Project Manager (software engineering) 2007
System Analyst (software engineering) 2005 to 2007
Program Analyst (software engineering) 2002 to 2004
BT Frontline Dalian, Dalian Liaoning, China

Professional Publication

Holsapple, C. W., & Yang, Z. (2013). Influence Structure and Inter-group Learning. *AMCIS 2013 Proceedings.*