ESSAYS ON DEVELOPING AND TESTING NEW MACHINE LEARNING MODELS FOR DYNAMIC MARKET SEGMENTATION AND PERSONALIZED PRODUCT RECOMMENDATION

A Dissertation

by

MILAD MOHAMMADI DARANI

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Venkatesh Shankar
Committee Members,	Shrihari Sridhar
	Bani Mallick
	James Caverlee
Head of Department,	David Griffith

August 2019

Major Subject: Marketing

ABSTRACT

This research comprises two essays on the development and application of machine learning models for marketing problems such as segmentation and personalized product recommendation. Prior market segmentation research has primarily focused on deriving customer segments based on demographics or brand choice data rather than on purchase sequences and attitudes. The first essay proposes a new machine learning approach that uses a mixture of hidden Markov models (MHMM) to cluster customers based on their purchase sequences of multiple items and predicts their cluster membership probabilities based on more than 200 attitude and demographic variables through a Lasso regression formulation. It shows the approach by estimating the models on a uniquely compiled dataset of transaction and attitudinal data in the salty snacks category from a large supermarket chain.

Accurate prediction of customers' next purchases is increasingly becoming important to retailers and manufacturers. Prediction of next purchases is a dynamic big data problem as it involves a large number of products and a huge base of customers, whose preferences may change over time. Current purchase prediction do not scale well and while machine learning topic models such as probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) model show promise, but they do not account for dynamics in customer preference. The second essay combines topic modeling and dynamic purchase modeling to propose a new machine learning approach, termed the Topic Hidden Markov Model (THMM) that is based on consumer preference theory. The model has several unique features that distinguishes it from alternative

models. First, the model can explicitly find latent segments of customers based on common preferences or motivations. Second, it captures heterogeneity by identifying idiosyncratic preferences. Third, the model incorporates dynamic patterns in customer preferences. Finally, it can learn correlated purchase patterns that exist in transactions of multiple items that are common in categories such as salty snacks and cereals. Using data from a supermarket retailer, the essay estimates the model, validates it, and benchmarks its product recommendations against those of alternative models. The model validation tests demonstrate the superior prediction accuracy of the new model and offer important implications for the theory and practice of marketing.

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance of my committee members. I would like to give my deepest gratitude to my committee chair, Dr. Venkatesh Shankar who encouraged me to work on this dissertation. He has been instrumental in providing me a sense of direction in my academic life. I would also like to show my gratitude to my committee members, Dr. Hari Sridhar, Dr. Bani Mallick, and Dr. James Caverlee for sharing their knowledge and time.

I would also like to thank the faculty, doctoral students, and staff of the marketing department. Dr. Alina Sorescu, thank you for being a wonderful PhD coordinator and for all of your help during the past five years. I appreciate our great group of marketing doctoral students for making this such an enjoyable experience.

Finally, I would like to thank my parents, Hamidreza and Tahereh, and my siblings, Farhad and Bahar who have always supported me unconditionally and have offered their help anytime and in any form I needed it. I am excited about what life is going to bring next.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Venky Shankar (chair of the committee) and Dr. Shrihari Sridhar of the Department of Marketing and Dr. Bani Mallick of the Department of Statistics and Dr. James Caverlee of the Department of Computer Science.

Funding Sources

This work was completed without any funding from the university or outside sources.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

We are witnessing a consensus across all the sectors of the society--from healthcare to politics to business--that data are the raw material of the future society. Digitalization has turned almost every business into an entity that collects and uses "big" data. Big data refer to data that exhibit five 'V's; volume, velocity, variety, veracity, and value. Volume refers to size, velocity to speed, variety to different types (e.g., structured, unstructured), veracity to integrity, and value to usefulness. Big data are particularly relevant to marketing as marketing deals with the gathering and analysis of information about customers and related decisions.

Marketers can collect so much information about their customers. Social media such as Facebook, Twitter, blogs, review websites, and forums are only a few examples of vast data sources available. Managers strive to answer questions like "what products should I recommend to each customer?" and "what is the most effective promotion for each group of customers." Managers need tools that can utilize large datasets that typically comprise millions of records and thousands of variables. Machine learning techniques are emerging as a promising set of tools that could help marketers derive insights from large datasets.

Machine learning involves systems, models, algorithms, and programs that learn patterns from existing data to make predictions about the future. Machine learning combines statistical techniques with computer science algorithms to help make decisions using data without the direct aid of humans.

In this research, I focus on two important marketing problems: market segmentation and personalized product recommendation. I build Machine learning models that capture the dynamic purchase behavior of customers to answer segmentation and recommendation problems.

With the explosion in data on consumer transactions, purchase sequences, and attitudes, companies are constantly seeking better methods to analyze such data and target the right customer segments with the right offerings and right marketing practices. Prior market segmentation research has primarily focused on deriving customer segments based on demographics or brand choice data rather than on purchase sequences and attitudes. The first essay proposes a new machine learning approach that uses a mixture of hidden Markov models (MHMM) to cluster customers based on their purchase sequences of multiple items and predicts their cluster membership probabilities based on more than 200 attitude and demographic variables through a Lasso regression variable selection model. It illustrates the approach by estimating the models on a uniquely compiled dataset of transaction and attitudinal data in the salty snacks category from a large supermarket chain.

Accurate prediction of customers' next purchases is increasingly becoming important to retailers and manufacturers. These predictions enable them recommend the right products to the right customers and develop effective and personalized recommender systems to enhance sales. Accurate predictions also help reduce excess inventory, prevent stockouts, and lower supply chain costs. Prediction of next purchases is a dynamic big data problem as it involves a large number of products and a huge base

of customers, whose preferences may change over time. Current purchase prediction models (e.g., collaborative filtering, stochastic models) do not scale well as it is managerially impractical to rely on product characteristics, limited customer characteristics, and data on sample of customers. Machine learning topic models such as probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) model show promise, but they do not account for dynamics in customer preference. The second essay combines topic modeling and dynamic purchase modeling to advance a new machine learning approach, termed the Topic Hidden Markov Model (THMM) that is based on consumer preference theory. The model has several unique features that distinguishes it from alternative models. First, the model can explicitly find latent segments of customers based on common preferences or motivations. Second, it captures heterogeneity by identifying idiosyncratic preferences. Third, the model incorporates dynamic patterns in customer preferences. Finally, it can learn correlated purchase patterns that exist in transactions of multiple items that are common in categories such as salty snacks and cereals. Using data from a supermarket retailer, the essay estimates the model, validates it, and benchmarks its product recommendations against those of alternative models. The model validation tests demonstrate the superior prediction accuracy of the new model and offer important implications for the theory and practice of marketing.

The remaining chapters are organized as follows. The next chapter offers an overview of the applications of machine learning techniques in marketing. Chapters III and IV capture the two essays and the last chapter summarizes and offers conclusions.

CHAPTER II

MACHINE LEARNING IN MARKETING

There is a long tradition of using a few machine learning (ML) methods in marketing. Every marketing researcher is likely to have used cluster analysis and logistic regression at some point in her career. However, except for such commonly used methods, ML has not been often used in marketing despite many ML techniques being well suited for addressing marketing problems.

Machine learning provides powerful and very efficient algorithms to tackle marketing problems. Consider, for example, the case of online marketing where decisions need to be made in real-time or close to real-time with a lot of data and variables also available. This is a scenario where machine learning techniques can shine.

In this chapter, I provide a brief overview of machine learning techniques and their applications in marketing. There is a vast literature on ML methods and their applications in various fields. Interested readers can refer to many books on the subject to get a deeper understanding of the literature (e.g., Hastie et al. 2009).

Machine learning techniques are methods – mostly developed by computer scientists – that learn patterns in historic data and make predictions about new unseen data. There is an overlap between ML and statistical methods and those that have been used in marketing. For example, regression analysis, and in particular, logistic regression, are among ML techniques that have been widely used in marketing. The biggest difference between the ML and classic statistical approaches is that in ML the

goal is to make accurate predictions, while in Statistics, the interest is in formal inferences about parameters and population.

ML techniques are usually categorized into three groups; supervised learning, unsupervised learning, and reinforcement learning. In the following paragraphs, I briefly introduce these categories and refer the reader to some papers in marketing that focused on those ML approaches.

Supervised Learning

Supervised learning is the most widely used approach in ML. In supervised learning, the objective is to predict the value of one or more outcome (dependent) variables using the value of a vector of input variables. The outcome variable could be continuous or discrete. When the outcome variable is discrete, the model is also called a classification model. In case of continuous outcome variable, the model is referred to as a prediction model. Artificial Neural Networks (ANN), Decision Trees, Support Vector Machines (SVM), and regression models are among the most Well-known supervised learning techniques.

In marketing, supervised models have been mainly used to predict customer behavior. The work by Cui and Curry (2005) is one of the early works on support vector machines in marketing. The authors compare the prediction accuracy of SVM with that of logit models in various marketing tasks. They show that SVM performs better and offer suggestions on how to apply SVM to solve marketing prediction problems. For an application of Artificial Neural Networks, see Thieme et al. (2000).

Drew and Ansari (2018) develop a nonparametric model (using Gaussian Priors) to predict customers purchase dynamics. Their model takes into account factors such as calendar time and events, customer life-time value, and interpurchase time. One advantage of the proposed model is that it provides model-based visualizations that incorporate hidden factors affecting customers spending propensity. The outputs are easy for managers to comprehend. The model can also capture the effect of interesting events like introducing a new product on customers spending.

Huang and Luo (2016) look into preference elicitation problem when dealing with complex products. They propose a framework that adaptively selects product profile questions for respondents based on his/her previous answers. They use a fuzzy SVM in building the adaptive learning mechanism. This SVM model predicts the response error on the fly and improves the survey design. Empirical and simulated studies prove the effectiveness of the proposed framework.

In the online space, there exists many recommender systems that are built based on machine learning algorithms. These algorithms must be both accurate and efficient. In most cases, the algorithms compare a focal customer's profile with other customers' profiles, and based on the similarities, decide which product might be of interest to the focal customer. The data used for training such models could be updated in real-time as new transactions occur, helping improve the accuracy of the models. Jacobs et al. (2016) identify an analogy between words frequency in a document and the number of times a customer purchase a product in her life-time. Because count data are involved, the authors use the idea of using a bag of words topic model to capture and predict customer

purchases. The authors use latent Dirichlet allocation (LDA) model to build their purchase prediction model.

The problem that Liu et al. (2016) tackle is pretty straightforward but also poses many practical challenges. How can we forecast customer behavior using online platform data? More specifically, the authors try to predict TV ratings using structured and unstructured data from Twitter, Google Trends, and IMDB (Internet Movie Data Base). The authors employ a dynamic panel data linear model as Well as some machine learning benchmark models to predict Nielson ratings of some TV programs. An interesting aspect of their model is that it uses MapReduce, a data reduction algorithm, for analyzing a big unstructured data set.

Rutz et al. (2017) build a model to predict the click performance of new ads. In the first step, they collect primary data on paired comparison of a training set of ads. Using the Elo method, the authors rank order the training ads. In the second step, they claim that the performance of an ad could be linked to its textual content and based on this notion they use various text mining techniques (e.g. bag of words and LDA) to extract textual features of ads. Then they use these features in a regularized regression called VANISH to predict the perceptual attributes and ultimately the performance of ads.

The work by Rutz et al. (2011) is another example of using a regularized regression model in marketing. The problem addressed here is measuring the indirect effect of paid search advertising. Indirect effect is defined as the conversion of customers' later visits to purchases at the website after they initially visit the website

through paid search advertising. The big challenge is that the data on website traffic is usually aggregated and if managers want to measure the impact of different keywords on the indirect effect, they face a "large p, small n" problem. The authors propose a hierarchical Bayesian elastic net model that effectively handles this problem. They find that indirect effect exists and is quite significant. They also identify some keywords that generate most of the indirect traffic.

Jaworska and Sydow (2008) use machine learning in online advertising and targeting context. Systems such as banner advertising as Well as search engine advertising use real-time auctions to sell ad space. Advertisers should decide in milliseconds how much an incoming customer is worth and how much money they want to invest to acquire the ad space. The task of predicting a customer's worth is a complex task and it is ideal for the application of machine learning models. Models need to predict whether a customer will click on the displayed ad, and if she clicked, whether it will result in a purchase.

Unsupervised Learning

Social network analysis, genes clustering and market segmentation are among the areas of most successful applications of unsupervised learning methods. In unsupervised learning, the training data consists of a set of only input vectors. Cluster analysis or clustering is the most common task addressed by unsupervised learning. In this task, the goal is to identify underlying structures in data. Clustering is a very well-known methodology in marketing. Various techniques of cluster analysis like hierarchical clustering, k-means clustering are already commonly used in marketing to identify

customer segments (see Wedel and Kamakura (2000) for an extensive review of cluster analysis in marketing).

Sometimes marketing models comprise both supervised and unsupervised learning tasks. Latent class analysis (Wedel and Kamakura 2000) illustrates how marketing researchers solve an unsupervised learning task while estimating model parameters in a supervised manner. Consider the latent class regression model. The goal is to find latent homogenous groups of customers (unsupervised learning). For every group, however, we need to train a logistic regression model that predicts customers' choices of brands given some independent variables (supervised learning). Thus, this method uses both supervised and unsupervised learning in one system.

Reinforcement Learning

Reinforcement Learning (RL) is generally used in applications such as robotics, control, finance, and inventory management. The common feature of these problems is a dynamic system that responds to decision maker's actions (as well as to its environment) by a change in its state. The goal is to learn an optimal policy or decision rule to maximize a function of rewards. This policy is a mapping between the states of the system and actions to be performed by the decision maker. A well-known example of reinforcement learning is the actions of a robot trying to keep its balance by controlling the motions of its parts.

RL allows agents to learn by exploring the action space and refining their behavior using only an evaluative feedback, referred to as the reward. In most cases, the agent's goal is to maximize its expected reward in long-run. As a result, the agent does

not just take into account the immediate reward, but it evaluates the consequences of its actions on the future. Reinforcement learning also includes dynamic programming and Markov decision processes.

RL models have attracted some attention in marketing. Schwartz et al. (2017) investigate the problem of sequential adaptive experiments in the context of online advertising. For a certain period, advertisers try different versions of an ad to identify the one that performs better. Then they allocate all the impressions to the high performing version of the ad. This process has two phases, exploration and exploitation. The authors use a reinforcement learning approach to find the optimal way of mixing exploration and exploitation, instead of switching between the two phases to maximize earnings of the firm.

Natural Language Processing (NLP)

Another prominent field of applied machine learning is natural language processing (NLP). Machine learning for NLP involves using machine learning algorithms to understand the meaning of text documents -- a task that could be challenging even for humans. These documents could include text from social media comments, online reviews, blog posts, and survey responses. Even financial, medical, legal and regulatory documents can be analyzed using NLP methods.

The role of ML techniques in natural language processing (NLP) is to improve and automate the underlying NLP techniques that turn unstructured text into useable information and insights. Machine learning for NLP and text analytics includes a set of statistical techniques for identifying parts of speech, entities, sentiments, and other

aspects of text. The underlying techniques can be either supervised machine learning or unsupervised machine learning methods.

In NLP using supervised machine learning, a set of text documents are tagged or labeled with examples of what the machine should search for and how it should interpret that aspect. These documents are used to train an ML model, which tests the predictions on untagged text. For example, with ML algorithms, researchers determine text sentiments. With the help of pre-tagged text sentiments (i.e., positive, negative, or neutral), the machine learns to understand how the occurrence of certain words determines the overall sentiment of a text. Unsupervised machine learning involves training a model without pre-tagging or annotating. Clustering similar documents in the same group and finding topics in a set of documents are examples of unsupervised techniques in NLP.

With the abundance of user generated content (UGC) from various online sources, marketing researchers have been studying the impact of such content on the sales of product. For example, Henning-Thurau et al. (2014) show that sentiments within UGC can predict sales. Chevalier and Mayzlin (2006) suggest that UGC can have a positive impact on consumer likelihood to buy products online. The application of other NLP techniques such as topic modeling holds promise for turning unstructured text data into data that could be used in traditional marketing models (e.g., Borah and Tellis 2016).

Performance Measurement and Related Issues

With the most statistical methods and machine learning techniques, as we add more independent variables or parameters to the model, we expect a better fit to the data. For instance, as we add independent variables to a regression model, the R-squared usually increases. Although we can explain more variance in the dependent variable, we cannot be sure that the model can make better predictions for new observations—a problem termed as overfitting. Similarly, many machine learning techniques are capable of using a large number of predictors to uncover linear and nonlinear relationships to fit an outcome variable in training data. However, when it comes to predicting the outcome for new cases, the model may not be generalizable.

To overcome the overfitting problem, researchers divide their data into training and test (hold-out) data. The allocation of observations is usually 75 to 80 percent for training and 20 to 25 percent for test. Researchers train and select those models that have the higher prediction accuracy in the test data. This is the approach that I take in developing and testing the models in the following two chapters.

CHAPTER III

DYNAMIC SEGMENTATION BY MULTIPLE ITEMS PURCHASES USING MACHINE LEARNING: LEVERAGING TRANSACTION AND ATTITUDINAL DATA

Introduction

Manufacturers and retailers are increasingly becoming customer-centric. With the explosion in data on consumer transactions, purchase sequences, and attitudes, companies are constantly seeking better methods to analyze such data and target the right customer segments with the right offerings and right marketing actions.

In any given product category (e.g., salty snacks), there exist multiple customer segments based on differences in customers' preferences and changes in those preferences over time. This dynamic preference change might manifest itself in consumer choice of different brands, purchase of different quantities (multiple items versus single items), or decisions to leverage promotions or deals. There could be groups of customers whose preferences may not change over time -- "static" customers. At the same time, there could be groups of customers whose preferences may change over time -- "dynamic" customers. It is important for managers to identify the static and dynamic customers as it would help managers design customized marketing mix campaigns.

With regard to brand choice and quantity determination, consumers could buy different combinations of multiple items on a purchase occasion. One segment of customers may primarily buy more quantities of one brand from one manufacturer (e.g., Lays potato chips from Frito-Lay); another segment may buy different brands from the

same manufacturer (e.g., Lays potato chips and Tostitos tortilla chips from Frito-Lay); and yet another segment may buy different brands from different manufacturers (e.g., Lays potato chips from Frito-Lay and Tops corn chips from Tops). Furthermore, customer membership in these segments may be dynamic. Customers who might have bought only one brand from one manufacturer in a category (belonging to the single brand-single manufacturer segment) in the past may buy more brands from the same manufacturer in the current or future periods, migrating to a different segment. Heterogeneity in purchase sequences across customers may determine such customer segmentation and a customer's demographics and attitudes may critically determine her segment membership.

Previous work in market segmentation has mainly used latent class modeling (LCM) for finding static segments of customers (Kamakura and Russell 1989). However, hidden Markov models have been used for finding dynamic segments of customers (Poulson 1990). My goal is to propose a market segmentation framework that uncovers both statics and dynamic segments. My proposed framework captures heterogeneity in purchase sequences or in dynamic behavior in a way that is not possible using a traditional Hidden Markov Model. My model is based on a mixture of Hidden Markov Models that allows us to detect dynamic groups that differ from other dynamic groups in the number of states among which customers transition.

In product categories with frequent purchases of multiple items, manufacturers need a better understanding of these customer segments to target the right customer segments with the right product and brand mix and the right promotions at the right time.

Retailers also need in-depth knowledge of these customer segments to market the right assortment of products and brands, including store brands to the right segments. Importantly, both manufacturers and retailers need to predict the probabilities of customer membership in these segments and customers' future purchases to better plan their marketing activities.

Considering the large number of manufacturers, brands, and sub-categories within a category, there exist several ways in which multiple item purchases can happen. As such, identifying segments based on multiple item purchases is a complex task but is invaluable to manufacturers and retailers.

To identify customer segments, manufacturers and retailers typically use consumer panel transaction data. Studies on market structure and market segmentation (e.g., Grover and Srinivasan 1987; Kamakura and Russell 1989) have focused on identifying latent segments based on brand choice within a category. These studies use mixture models in which customers are assumed to remain in one of unobserved segments that are recovered by latent class analysis.

Prior segmentation studies have important limitations with regard to analysis of multiple items. These studies do not consider quantity or multiple item purchases as criteria for segmentation and are static—that is, segment characteristics and composition do not change over time. As outlined earlier, customers could belong to different segments based on the combination of multiple items they purchase and their segment membership could be dynamic. Moreover, segmentation and assignment of customers to segments in prior models are based purely on behavioral (transaction) data. However, a

customer's attitudes can critically influence her membership in a segment (Wedel and Steenkamp 1991). A customer may be in one segment based on past purchases, but her attitude toward purchased brands and items may affect her future purchases of multiple items, potentially migrating her to a new segment. Therefore, modern manufacturers and retailers need to better predict a customer's future segment membership and purchases based on the sequence of the customer's past transactions as well as her attitudes. To classify customers into segments and predict each customer's membership in these segments and their future purchases, today's managers need models that can learn from all available data, that is, machine learning models that utilize both transaction and attitudinal data.

In this essay, I propose a novel machine learning approach to classify customers into different segments based on purchases of multiple items and to predict future customer membership in segments and purchases using both transaction and attitudinal data. My approach comprises a mixture of hidden Markov models (MHMM) based on sequence clustering using transaction data and a Lasso (Least Absolute Shrinkage and Selection Operator) regression variable selection model for predicting customers' segment membership using attitudinal data. I apply my approach to data on salty snacks purchases. I estimate the model using a training sample and validate it on a test sample. I derive valuable implications for both retailers and manufacturers.

This essay makes important contributions to the marketing literature. First, I propose a new machine learning approach to segment customers to enhance multiple item purchases that is important for manufacturer and retailer decisions. My approach

allows for segment formation to be dynamic. Second, my approach leverages both behavioral and attitudinal data to offer an in-depth understanding of customer attitude and behavior. Third, this model and predictions offer rich insights into future purchase sequences of customers. Finally, the proposed model and analysis provide key actionable implications for both manufacturers and retailers.

Related Literature

Wedel and Kamakura (2000) categorize the market segmentation literature into two groups: predictive methods and descriptive methods. Predictive methods look into the relationships between a set of variables (i.e., independent variables) and one or more outcome variables. Descriptive methods, however, do not distinguish between independent and outcome variables. Predictive methods allow decision makers to predict customer behavior and the profitability of a marketing mix based on customer demographics or past behavior. Descriptive methods help profile customer segments. However, this simplistic categorization does not fully capture the computational complexity of the market segmentation problem and the diversity and abundance of market segmentation techniques.

Prior research also suggests different ways to categorize market segmentation approaches (Wedel and Kistemaker 1989). One such way is based on type of applications. Techniques such as clustering and conjoint analysis can be used to segment based on application domains (e.g., customer retention, price discrimination, DeSarbo et al. 2008), application objectives (e.g., to explain the differences in customer choice behavior or maximize the effectiveness of market resources allocation among segments,

Mahajan and Jain 1978), and application data (e.g., product usage, demographic data, Kim et al. 2006).

Liu et al. (2012) take a more theoretical stance on categorizing the market segmentation literature. They view the literature along three dimensions, relationship among segmentation bases, data measurement, and the solution techniques for multi-objective optimization. Consistent with Wedel and Kamakura (2000), they classify segmentation methods as descriptive or predictive approaches depending on whether the method distinguishes between independent and dependent variables or not. Descriptive methods include clustering-based methods such as the variations of K-means (Chiu et al. 2009) or p-median (Klastorin 1985) methods. Some predictive methods for market segmentation are cluster-wise regression model (Späth 1979), cluster-wise logistic model, and mixture regression model.

Market segmentation methods can also be classified as discriminative (or distance/similarity-based) and generative approaches according to the computational assumption about the data distribution. For example, Euclidean distance is a common similarity measure for a pair of data points. However, when the data distribution is known, the results derived from generative models (e.g., finite mixture model) are usually more interpretable and allow for statistical inferences. Liu et al. (2012) also suggest that the market segmentation problem is a multi-criteria problem and that existing market segmentation methods can be classified into multi-stage, transformation, and multi-objective optimization approaches.

As outlined earlier, market segmentation methods have been used in various application areas. One that is closely related to my work is the stream of research that infers market structure from consumer preferences or choice data. From a marketing standpoint, internal market structure analysis refers to the study of customers' brand preferences based on brand attributes and customer preferences for those attributes (Elrod 1991). For a summary of market structure analysis, see Elrod and Keane (1995). Brand switching data have been used for segmenting consumers into loyal and switching segments (e.g., Grover and Srinivasan 1987). Some studies have analyzed purchase frequencies (e.g., Elrod 1988) or disaggregated panel data (e.g., Elrod and Keane 1995) to understand market structure.

Some studies combine market structure analysis and market segmentation. DeSarbo et al. (1993) use a stochastic tree-unfolding method to determine market structure and market segments at the same time. They identify product-market hierarchy using paired-comparison preference judgments. Similarly, Wedel and Steenkamp (1991) propose a generalized fuzzy cluster-wise regression approach that simultaneously estimates fuzzy market structures and fuzzy segments based on customer benefits.

These segmentation studies typically do not incorporate information about the complete sequence of purchases. Some studies that include information about purchase histories typically do this in a limited way. For instance, Erdem (1996) estimates a dynamic model using state dependence (based on the previous purchase) and concludes that not accounting for state dependence may lead to biased results. Grover and Srinivasan (1987) estimate a latent class model to derive segments of loyal and

switching consumers but use information from only the aggregate switching matrix. They include two randomly selected purchase occasions based on the assumption that purchase behavior is zero-order. Kamakura and Russell (1989) develop a mixture regression model to segment consumers who are homogeneous in price and sales promotion elasticities. Although they use household scanner panel data to estimate the household-level (not market-level) elasticity, the likelihood of the observed choice histories belonging to specific segments are independent of the sequential relationships between the choices of a consumer.

Multiple item purchases have received some attention in the marketing literature in the form of purchases of a bundle or collection of items. Farquhar and Rao (1976) propose a model that maximizes either the average utility in the bundle or its variance to capture the "heterogeneity is better" idea in the bundle. Chung and Rao (2003) and Harlam and Lodish (1995) extend this model to predict the purchase of bundles.

Some structural models examine multiple purchases in an indirect manner. Even within a product category of close substitutes such as soft drinks or yogurt, consumers may select an assortment of products on one shopping trip. A model of multiple discreteness explains the choice of a single-category assortment as the outcome of a utility process in which consumers simultaneously choose items and quantities. Dubé (2004) argues that consumers anticipate future consumption occasions and accordingly maximize a utility function defined across these occasions. The consumer's maximization problem yields a mixture of interior (positive quantity) and corner (no consumption) solutions, in other words, a bundle of products. Kim et al. (2007) develop

an alternative approach to the multiple discreteness problem in which product characteristics are projected onto the utility space.

Prior segmentation approaches are not dynamic in that they do not typically take into account the sequence of purchases. Moreover, there is scant research on segmentation of customers based on multiple items purchases. As argued earlier, a segmentation approach based on sequence of purchases multiple items is invaluable to both manufacturers and retailers. My study develops such a segmentation approach. A comparison of my study relative to selected related literature appears in Table 1. Unlike related approaches, my approach comprises all the three key elements, market segmentation, dynamic treatment, and multiple item purchases and uses both household panel data and attitudinal data.

Table 1 A Comparison of my Study with Selected Related Literature

Paper	Data	Market Segmentation	Dynamic Treatment	Multiple Items
Grover and Srinivasan (1987)	Household scanner panel data	Yes	No	No
Kamakura and Russell (1989)	Household scanner panel data	Yes	No	No
Wedel and Steenkamp (1991)	Survey data of consumer preferences	Yes	No	No
Elrod and Keane (1995)	Household scanner panel data	Yes	Yes	No
Harlam and Lodish (1995)	Household scanner panel data	No	No	Yes
Erdem (1996)	Household scanner panel data	No	Yes	No
Dubé (2004)	Household scanner panel data	No	No	Yes
Kim, Allenby, and Rossi (2007)	Household scanner panel data	No	No	Yes
My Paper (2018)	Household scanner panel data, Attitudinal data	Yes	Yes	Yes

A Clustering Model of Multiple Items Purchases

I develop a multiple item purchases model that has both descriptive and predictive uses. I start with the descriptive part of the model that analyzes and identifies clusters of multiple item purchase sequences. I follow it with a brief discussion of the predictive part but defer the full discussion on this part to the results section.

Transaction Clustering

The first step in my framework is identification of types of transactions. I focus on four dimensions of transactions and identify clusters of transactions that are homogenous in these dimensions: (1) brands purchased, (2) quantity purchased (single or multiple items), (3) quantity of items purchased within a multiple item transaction, and (4) use of deal (whether the purchased item was on promotion or a coupon was redeemed during purchase)

To perform such clustering, I examine methods from the computer science literature. One class of methods relies on finding association rules (Han and Fu; 1999). However, most of these methods cannot handle scenarios where consumers purchase more than one unit of an item on a shopping occasion. The association rules in the literature cannot also be used for clustering transactions into groups of similar patterns of multiple item transactions. Therefore, I propose the following heuristic procedure to deeply understand and cluster transactions.

First, we need a new format for representing the transactions. We would like to work with purchase quantities and the use of deals. Assume we have $l = 1, \ldots, L$ different brands in a transaction and N_l is the name of each brand. Furthermore, assume

that the purchased quantity of each brand is q_l. For every transaction, I create a label that has $\mathrm{L}+2$ components. For every brand, I create a component $\mathrm{N}_l = \begin{cases} 1 \; if \; q_l = 1 \\ 2 \; if \; q_l > 1 \end{cases}$. I also define component $Ty = \begin{cases} S \; if \; \sum q_l = 1 \\ M \; if \; \sum q_l > 1 \end{cases}$ that shows whether there are multiple items in the transaction, and component $= \begin{cases} D \; if \; a \; deal \; used \\ ND \; if \; no \; deal \; used \end{cases}$. Therefore, I transform every transaction to a vector $(N_1, \ldots, N_L, Ty, P)$.

The new format of data representation allows us to cluster transactions based on multiple bases. Most of the clustering techniques work using similarity (or dissimilarity) between data points. Therefore, I need to use a similarity measure for the transactions to be able to cluster them. The Jaccard index is a suitable metric for measuring the similarity between two sets of items and is defined as the size of intersection divided by the size of the union of the two sets.

In the clustering part, I use an implementation of the k-medoids algorithm called Partition Around Medoids (PAM) (Kaufman and Rousseeuw 1987). The PAM algorithm selects a set of S items as the clusters' medoids to minimize the average dissimilarity between objects and their closest selected object. I specify the number of clusters. To determine the number of clusters, I try different numbers of clusters and choose the one that generates a clustering solution with an adequately high average silhouette width[1].

[1] For more details about the silhouette measure see the results section of this chapter.

To improve the prediction accuracy of the proposed model, we need make better predictions about transactions that belong to a certain cluster. To do accomplish this, I look for the most frequent purchase patterns in every transaction. I call a set of items an itemset. To inspect clusters and have a better understanding of the type of transactions they contain, I search for "frequent itemsets" in every cluster. An itemset with k items is called a k-itemset. The $support$ of an itemset X, denoted by $\sigma(X)$, is the number of transactions in which it occurs as a subset. An itemset is *frequent* if its support is more than a user-specified minimum support. The set of frequent k-itemsets is denoted by $\mathcal{F}(k)$. As I will subsequently show in the results section, identifying frequent itemsets is an effective way to summarize, show the content of clusters of transactions, and make predictions about them.

Purchase Sequence Clustering

In the previous section, I proposed a framework for clustering all the transactions in the dataset. However, customers might show different buying habits over time. For example, in the context of multiple item purchases, a customer might purchase from two different brands in a trip with the intention of trying a new brand in addition to his/her favorite brand. To capture the patterns that happen over time, I need to cluster customers based on their sequence of purchases. I use a mixture of hidden Markov models (MHMM) to cluster purchase sequences. I replace each transaction with the name of the cluster to which it belongs. Therefore, for each customer in the database, I have a sequence of multiple item purchase clusters.

In the context of Hidden Markov models, sequence data include observed states that are considered to be probabilistic functions of hidden states. The hidden states cannot be observed directly but can be inferred through the sequence(s) of observations since they emit the observations on varying probabilities. A discrete first order hidden Markov model for a single sequence is determined by the following:

Observed state sequence or observed transaction cluster sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ with observed transactions $m \in \{1, \ldots, M\}$, and T shows time index.

Hidden state sequence $\mathbf{z} = (z_1, z_2, \ldots, z_T)$ with hidden states $s \in \{1, \ldots, S\}$.

Transition matrix $A = \{a_{sr}\}$ of size $S \times S$, where a_{sr} is the probability of moving from the hidden state s at time $t-1$ to the hidden state r at time t:

(1) $$a_{sr} = P(z_t = r | z_{t-1} = s); \quad s, r \in \{1, \ldots, S\}$$

I consider only homogeneous HMMs in which the transition probabilities are constant over time.

Emission matrix $B = b_s(m)$ of size $S \times M$, where $b_s(m)$ is the probability of the hidden state s emitting the observed transaction m:

(2) $$b_s(m) = P(y_t = m | z_t = s); \quad s \in \{1, \ldots, S\}, m \in \{1, \ldots, M\}$$

Initial probability vector $\pi = \{\pi_s\}$ of length S, where π_s is the probability of starting from the hidden state s:

(3) $$\pi_s = P(z_1 = s); \quad s \in \{1, \ldots, S\}$$

The (first order) Markov assumption states that the hidden state transition probability at time t only depends on the hidden state at the previous time point $(t-1)$:

(4) $$P(z_t | z_{t-1}, \ldots, z_1) = P(z_t | z_{t-1})$$

Furthermore, the observation at time t is only dependent on the current hidden state, and not on previous hidden states or observations:

$$P(y_t|y_{t-1},\dots,y_1,z_t,\dots,z_1) = P(y_t|z_t) \quad (5)$$

For a more detailed description of hidden Markov models, see Rabiner (1989), MacDonald and Zucchini (1997), and Netzer et al. (2008).

Now assume that we have N customers in the database. Therefore, instead of one observed sequence $\mathbf{y}$, we have N sequences as $Y = (\mathbf{y}_1,\dots,\mathbf{y}_N)^T$, where the observations $\mathbf{y}_i = (y_{i1},\dots,y_{iT})$ of each customer i take values in the observed transaction space.

Estimation

I estimate the unknown transition, emission and initial probabilities via maximum likelihood. The log-likelihood for multiple sequences is written as:

$$\log L = \sum_{i=1}^{N} \log P(y_i|\mathcal{M}) \quad (6)$$

Where $\mathcal{M}$ describes the hidden Markov model and its parameters $\{\pi, A, B\}$.

Clustering with a Mixture of Hidden Markov Models

Assume that I have a set of models $\mathcal{M} = \{\mathcal{M}^1,\dots,\mathcal{M}^K\}$, where $\mathcal{M}^k = \{\pi^k, A^k, B^k\}$ for $k = 1,\dots,K$. For each customer y_i, denote $P(\mathcal{M}^k) = w_k$ as the prior probability that the observation sequences of customer i belongs to the submodel/cluster $\mathcal{M}^k$. Now the log-likelihood is extended from Equation 6 as:

$$\begin{aligned} \log L &= \sum_{i=1}^{N} \log P(y_i|\mathcal{M}) \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(\mathcal{M}^k) \sum_{allz} P(y_i|z,\mathcal{M}^k) P(z|\mathcal{M}^k)\right] \\ &= \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \sum_{allz} \pi_{z_1}^k b_{z_1}^k(y_{i1}) \prod_{t=2}^{T} [a_{z_{t-1}z_t}^k b_{z_t}^k(y_{it})]\right] \end{aligned} \quad (7)$$

The posterior cluster probabilities $P(\mathcal{M}^k|Y_i, \mathbf{x}_i)$, where $\mathbf{x}_i$ is a customer's covariate values, are obtained as:

$$P(\mathcal{M}^k|y_i, \mathbf{x}_i) = \frac{P(y_i|\mathcal{M}^k, \mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{P(y_i|\mathbf{x}_i)} = \frac{P(y_i|\mathcal{M}^k, \mathbf{x}_i)P(\mathcal{M}^k|\mathbf{x}_i)}{\sum_{j=1}^{K} P(y_i|\mathcal{M}^j, \mathbf{x}_i)P(\mathcal{M}^j|\mathbf{x}_i)} \quad (9)$$

Data and Results

Data

I compiled a unique dataset comprising behavioral (household scanner panel) data and attitudinal (survey) data about salty snacks from a New York state grocery chain. The dataset includes purchase histories of 10000 households and results of a survey of households' shopping habits of 2500 of these households. The survey comprises several questions about customer demographics and attitudes toward salty snacks shopping. After dropping variables with missing values, my data contain 212 attitude and demographic variables, such as perception of snacks and income. Table 2 shows some descriptive statistics of customer purchase behavior and attitudes in the salty snacks category in my data. Specifically, it summarizes the information about the items observed in every transaction.

I began my analysis by selecting all the transactions containing one to five items from the salty snacks category for the 2,500 households.

Results on Clustering Transactions

Because running the machine learning clustering algorithm is time-intensive, I performed some preliminary analyses to reduce the sample size to a meaningful set. The preliminary analyses showed that the number of transactions with more than five items is

very small and negligible compared to the number of transactions with one to five salty snack items. Therefore, I dropped the transactions (and corresponding customers) with more than five items. I also excluded the last transaction of every customer from the training data for testing the accuracy of the model. I ended up with 1,759 customers and 46,859 transactions. To keep the runtime reasonable, I drew a random sample of 3,000 transactions from this set of transactions. A subsequent robustness check showed that this smaller sample does not affect the final results of clustering.

Table 2 Summary Statistics of Purchase Behavior and Attitudinal Data

		Min	Max	Mean	Standard Deviation
Transaction Level	Number of items purchased in the transaction	1	16	1.93	1.17
	Number of distinct UPCs purchased	1	11	1.53	0.85
	Number of distinct brands purchased	1	8	1.34	0.64
	Number of distinct manufacturers	1	7	1.24	0.52
Household Level	Number of transactions	1	135	13.7	14.38
	Number of distinct UPCs purchased	1	73	11.62	10.89
	Number of distinct brands	1	24	6.15	4.38
	Number of distinct manufacturers	1	15	4.17	2.7
	Selected Attitudinal Variables				
	I am willing to trade off the lower quality in store brand salty snacks for their lower price.	1	7	3.35	1.73
	I don't believe that "2 for 1" salty snack deals save you much money.	1	7	5.08	1.61
	Purchasing store brands of salty snacks is riskier because they offer less value for the money than national brands.	1	7	3.21	1.58
	When I buy salty snacks, I first think of the product type before deciding on a particular brand.	1	7	4.9	1.75
	I love shopping.	1	10	5.7	2.6

In the next step, I calculated the dissimilarity matrix that contains pairwise dissimilarities between transactions. As outlined earlier, I used the Jaccard index for calculating dissimilarities. Using the PAM techniques, I clustered the sample transactions.

To find a good clustering solution, I calculated a silhouette measure for each item. This measure shows how similar each item is to its cluster and how different it is from neighboring clusters. The solution with 10 clusters of transactions had the best average silhouette width. Figure 1 depicts the quality measures for this clustering. Note that silhouette widths close to one show a very high quality of clustering solution.

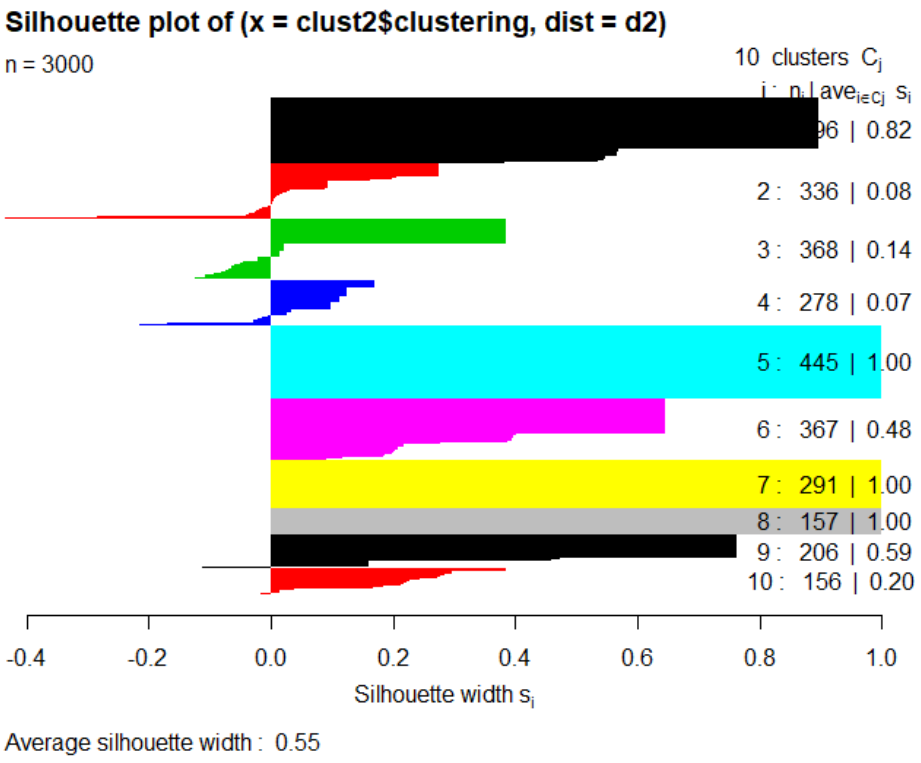

Figure 1 Silhouette Width Graph for Transaction Clustering with 10 Clusters

To investigate the content of each cluster, I need a summarization technique. I use the Eclat algorithm to find the most frequent items in each cluster. The Eclat is a fast algorithm for finding frequent itemsets[2]. Since my clusters are not really big both in number of transactions and number of items, even inefficient heuristics would be enough to quickly find frequent itemsets. However, with large clusters, we need to use more efficient techniques like Eclat. Figure 2 shows the content of the sixth cluster of the clustering solution.

In Figure 2, we see relative frequencies of items in the cluster. All the frequent items with the minimum support of more than 0.02 are shown in the graph. For instance, "FRTO=2" shows a very high frequency of multiple items of FritoLay in this cluster. The line in the graph shows the overall frequency of items in all clusters.

I obtained the ten clusters depicted in Table 3. The third column shows the most frequent patterns with size one in every cluster with information about support and count of every item.

[2] Identifying frequent itemsets is a non-trivial task and a huge body of literature in computer science is devoted to this problem and similar problems (see Han and Fu; 1999 for details).

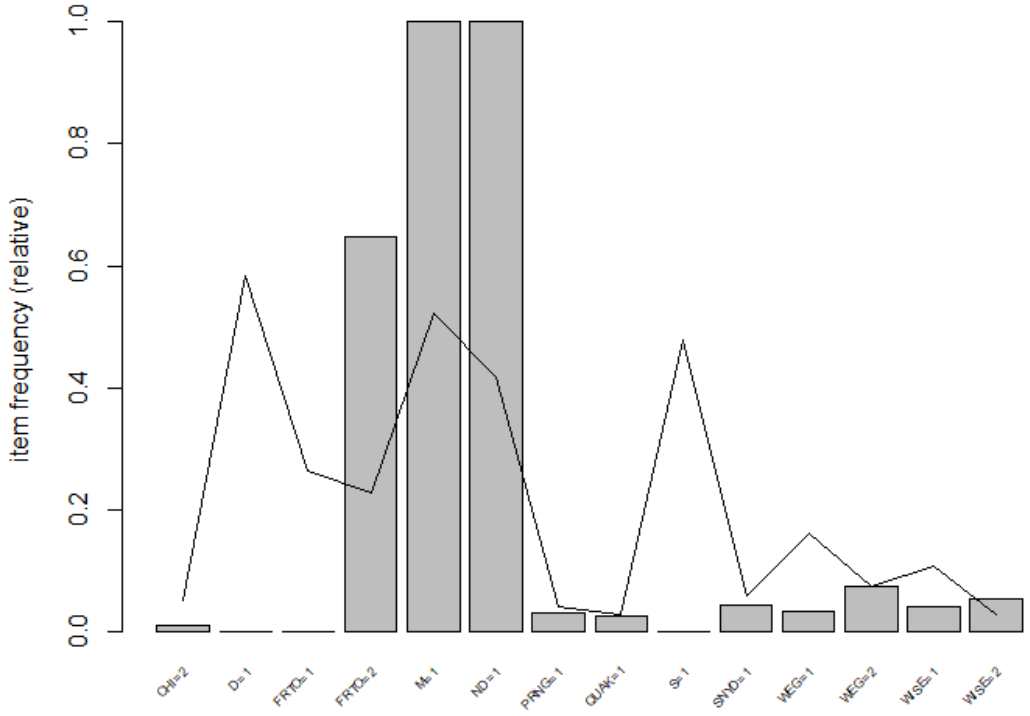

Figure 2 Item Frequency Distribution for Cluster 6 from Transaction Clustering

Results on Clustering Purchase Sequences

We can convert the sequences of multiple item purchases into sequences of clusters to which every transaction belongs. This summarization makes the task of analyzing sequences more tractable. Note that we have sequences of purchase occasions and we observe different sequence lengths for different customers.

Table 3 Clusters of Multiple Item Purchases

Number	Size	Frequent patterns
1	7694	items support count [1] {D=1} 1.00000000 7694 [2] {M=1} 1.00000000 7694 [3] {FRTO=2} 0.84793345 6524 [4] {WEG=1} 0.06654536 512 [5] {WISE=1} 0.05224851 402
2	3361	items support count [1] {M=1} 1.00000000 3361 [2] {D=1} 0.99196668 3334 [3] {CHI=2} 0.62600417 2104 [4] {FRTO=1} 0.09520976 320 [5] {WEG=1} 0.09163939 308 [6] {WISE=1} 0.06932461 233 [7] {WISE=2} 0.08687891 292
3	3436	items support count [1] {D=1} 1.00000000 3436 [2] {S=1} 0.94761350 3256 [3] {WISE=1} 0.74796275 2570 [4] {SNYD=1} 0.07712456 265 [5] {M=1} 0.05238650 180
4	2337	items support count [1] {ND=1} 1.00000000 2337 [2] {S=1} 0.96106119 2246 [3] {WISE=1} 0.32691485 764 [4] {SNYD=1} 0.17372700 406 [5] {PRNG=1} 0.12965340 303 [6] {QUAK=1} 0.09456568 221 [7] {WEG=1} 0.08472401 198 [8] {NMKT=1} 0.05562687 130
5	8134	items support count [1] {ND=1} 1.000000 8134 [2] {S=1} 1.000000 8134 [3] {FRTO=1} 0.805016 6548
6	6106	items support count [1] {M=1} 1.00000000 6106 [2] {ND=1} 1.00000000 6106 [3] {FRTO=2} 0.64837864 3959 [4] {WEG=2} 0.07533574 460 [5] {PRNG=2} 0.06092368 372 [6] {WISE=2} 0.05470029 334
7	5720	items support count [1] {D=1} 0.9667832 5530 [2] {S=1} 0.9729021 5565 [3] {IG=1} 0.8466783 4843
8	3412	Items support count [1] {D=1} 1.00000000 3412 [2] {S=1} 0.94724502 3232 [3] {FRTO=1} 0.75527550 2577 [4] {SNYD=1} 0.07649472 261 [5] {M=1} 0.05275498 180

Table 3 Continued

Number	Size	Frequent patterns
9	3960	items support count [1] {M=1} 1.00000000 3960 [2] {D=1} 0.90909091 3600 [3] {WEG=2} 0.69823232 2765 [4] {FRTO=1} 0.07525253 298 [5] {ND=1} 0.09090909 360 [6] {WISE=1} 0.06565657 260 [7] {WISE=2} 0.06388889 253
10	2699	items support count [1] {M=1} 1.00000000 2699 [2] {FRTO=1} 0.88847721 2398 [3] {ND=1} 0.86624676 2338 [4] {WEG=1} 0.39903668 1077 [5] {WISE=1} 0.17932568 484 [6] {D=1} 0.13375324 361 [7] {SNYD=1} 0.12412004 335 [8] {PRNG=1} 0.11411634 308 [9] {QUAK=1}0.05854020 158 [10]{WEG=2} 0.05446462 147

Before estimating the hidden Markov models, I deal with varying lengths of sequences and missing observations. When an observation is missing, we gain no additional information regarding hidden states. I handle sequences of varying lengths by setting missing values before and/or after the observed purchase.

To estimate a MHMM model, I need to specify the number of clusters or models and the number of hidden states for each model. I estimated different combinations of number of models and clusters and compared them based on prediction accuracy of the last transactions of 1,759 customers. I also trained a latent class model as well as a single hidden Markov model as benchmarks. Table 4 shows the prediction accuracy of the best fitting models. I compare models on three prediction tasks. I predict brands purchased, transaction type (multiple item versus single item), and use of deals in the transaction.

For the transaction type predicted, I examine the frequent items from Table 3 to predict actual behavior.

From Table 4, the mixture of HMMs is more accurate in predicting brands purchased and the transaction type than the other two benchmark models. This higher accuracy means that the mixture model can explain revenues up to a three percent greater accuracy than a single HMM can. However, the single HMM performs better in predicting the use of deals in the transaction, and there is no difference between the MHMM model and the latent class model in this aspect. The solution of the MHMM includes nine hidden Markov models, the single hidden Markov has three hidden states, and the latent class model has seven segments.

Table 4 Prediction Accuracy of MHMM with Benchmark Models

	Mixture of HMM	Single HMM	LCM (Latent Class)	Random
Brands purchased	62.1%	60%	59%	20%
Purchase of multiple items	62%	56.6%	60%	50%
Using deals	81.7%	85.1%	82%	48%

A further examination of MHMM shows that the transition probabilities among the hidden states in some HMMs are small (less than 0.01). Thus, these states behave like static segments. Therefore, I can summarize the final segmentation solution as in Table 5 with four static segments, two HMMs with two hidden states, and five HMMs with three hidden states in the mixture. The last column in the table shows the most likely transaction type in every state.

Cluster 1: This cluster is a static cluster. The most probable transaction type made by the customers in this segment is transaction type 9, which means that this customer group tends to purchase multiple items of salty snacks. They are very likely to buy more than one item of store brand and use a deal or promotion in their transactions.

Cluster 2: This cluster is another static cluster. The dominant behavior is transaction type 7, suggesting these customers prefer single item of a store brand on a deal.

Cluster 3: This cluster is also a static cluster. Customers in this cluster tend to purchase Frito Lay brands without using any promotions or deals. They are almost equally likely to purchasing single items and multiple items.

Cluster 4: Cluster 4 is another static cluster. Customers in this cluster tend to purchase a single item in a transaction. FritoLay is the most popular choice that is often purchased without using any promotions.

Cluster 5: This cluster is a dynamic cluster with two states. Customers exhibit two different types of purchase behavior over time. Customers in this cluster exhibit both types of behavior that static customers in Clusters 3 and 4 display. In other words, these customers are interested in FritoLay and they don't use promotions. However, they buy multiple items during some purchase occasions and only one item during other occasions. The transition probability from purchasing a single item to purchasing multiple items is 0.3 and the transition probability from purchasing multiple items to a single item is 0.16. This result shows that multiple item purchases are stickier than single item purchases.

Cluster 6: This is another dynamic segment with two states. The common behavior in this segment is that customers purchase multiple items on promotion. The dynamic behavior, however, stems from brand choice. Customers switch between FritoLay and another national brand (CHI). One reason is that these customers are price sensitive and would buy another brand if it is offered with a more valuable promotion.

Cluster 7: This is a dynamic cluster with three hidden states with three types of dominant behavior. In the first state, the consumers purchase multiple items of FritoLay on promotion (transaction type 1). In the second state, they tend to make transaction type 7, where consumers buy a single item of store brand under promotion. In the third state, consumers purchase multiple items of FritoLay with no deal (transaction type 6). To summarize, this segment purchases multiple items of FritoLay (with or without promotions). But when it comes to the store brand, consumers purchase only a single item.

Cluster 8: This group is also a dynamic group with three states. The common behavior is purchasing single items from national brands. This group of consumers show little interest in the store brand. The dynamic variable in this segment is deal use. There is also some variation in the choice of national brands.

Cluster 9: This is yet another dynamic group with three hidden states. This is another segment that store brand managers may want to target. In the first state, the consumers purchase single item of FritoLay not on promotion (transaction type 5). In the second state, consumers purchase multiple items of FritoLay on promotion (transaction type 1). In the third state, consumers tend to buy either a single item or multiple items of

store brand on promotion. This group is similar to Cluster 7 with two differences. First, consumers in Cluster 9 are less susceptible to transitioning to another state. In other words, the diagonal elements in transition probability matrix for Cluster 9 are quite large. Second, Cluster 9 consumers are more price sensitive and buy multiple items only on promotion. To summarize, these consumers are loyal to their brand but buy multiple items only under valuable promotions.

Cluster 10: This segment is another dynamic segment with three states. The first state represents purchases of single item of the store brand under promotion. The second state is dominated by single item purchases of FritoLay with no promotion. In the third state, consumers tend to buy either a single item or multiple items of the store brand under promotion. This group of consumers is loyal to the store brand and only occasionally purchase a FritoLay product.

Cluster 11: This dynamic segment also has three states. This is the most variety seeking group of customers in the data. The consumers in this segment exhibit three different types of behavior characterized by their choice of brands: store brand, FritoLay, and smaller national brands. These consumers are price sensitive. They typically purchase during promotions or deals.

Predictive Model of Segment Membership

To predict segment membership of consumers, I tried adding the attitudinal and demographic data to the mixture hidden Markov model. In such a model, the cluster membership probabilities are functions of covariates. However, simply adding covariates to the MHMM imposed additional constraints, leading to a lower model fit.

Table 5 Mixture of HMM's Segmentation Solution

	Type	Number of States	Most Probable Transaction Type (Relative Frequency)
1	Static	1	9(41%)
2	Static	1	7(55%)
3	Static	1	5(38%)
4	Static	1	5(65%)
5	Dynamic	2	5(41%), 6(22%)
6	Dynamic	2	1(34%), 2(38%)
7	Dynamic	3	1(18%), 7(20%), 6(37%)
8	Dynamic	3	3(23%), 3(34%), 5(41%)
9	Dynamic	3	5(36%), 1(40%), 7(36%)
10	Dynamic	3	7(34%), 5(42%), 7(31%)
11	Dynamic	3	7(19%), 1(17%), 3(54%)

I use two alternative prediction methods, namely, neural network and Lasso regression (Tibshirani 1996)[3]. Because my target variable, cluster membership probability, is a continuous variable, it rules out the use of classification methods such as decision trees. For the prediction exercise, we are restricted to the subsample of 474 customers who were randomly chosen for the survey. I divided them into training and test samples, assigning 80% of the observations to the training sample. I trained the best

[3] See Appendix A for more details about Lasso regression.

mixture of HMMs and predicted the first transaction of consumers in the holdout/test sample.

As benchmark models, I trained multiple neural networks with different numbers of units in the hidden layer to predict each cluster's membership probability. The number of units in the hidden layer does not have a huge influence on the model fit in my case. It only increased the time the network needed to converge. Therefore, I report only the results with five units in the hidden layer. I set the number of maximum iterations in case the model did not converge in a reasonable amount of time to 100.

For the Lasso regression model, I had to set the tuning parameter λ that controls the penalty in the training sample. The higher the value of λ, the more restrictive the magnitude of the coefficients. I used cross validation to choose the optimal λ. Figure 3 shows the Multinomial deviance for different values of λ. I selected the optimal values of λ based on cross-validation deviance to train the models.

After training both the models, neural network and Lasso regression, I used the test or prediction sample to compare the accuracy of the two techniques. I achieved 55 percent prediction accuracy with the Lasso model and 54 percent with neural net, which is not significantly lower than that for the Lasso model.

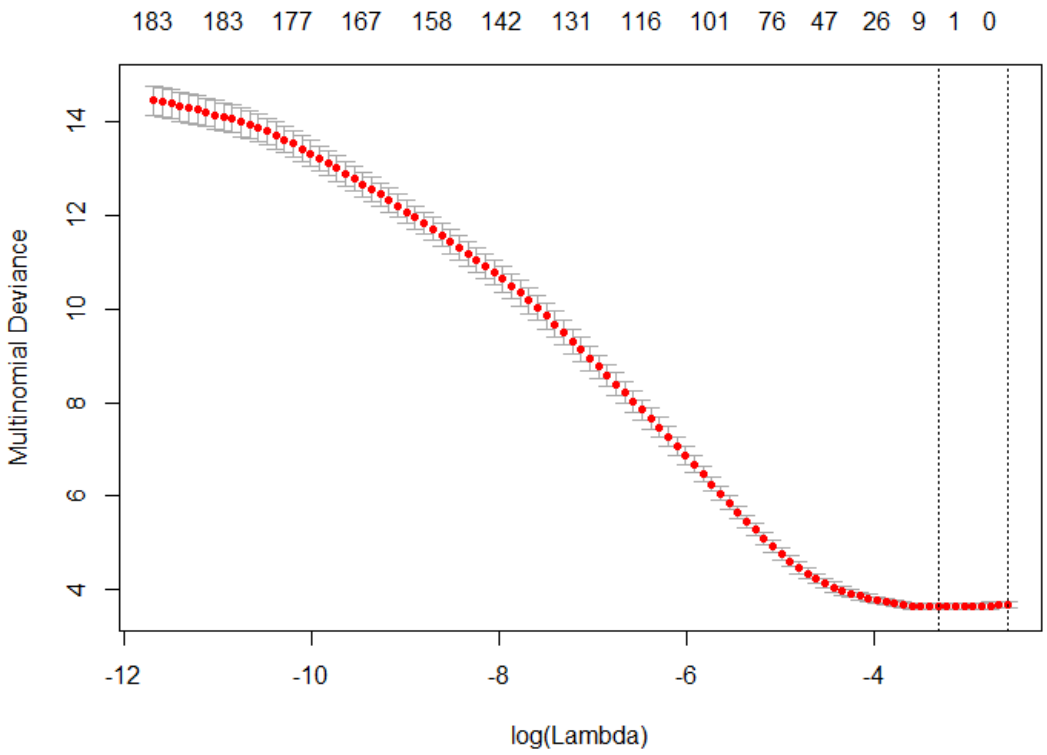

Figure 3 Multinomial Deviance vs. Tuning Parameter λ in Lasso Regression

The biggest advantage of Lasso regression over neural network is that the output of Lasso regression is interpretable, while neural network is a black box. We can interpret the Lasso coefficients the same way as we do for OLS regression model coefficients. However, the coefficients of the Lasso regression do not carry information on statistical significance. Table 6 reports some of the variables that survived the Lasso regression, i.e., variables with non-zero coefficients in the linear model. Thirty seven variables effectively predict segment memberships.

Table 6 Lasso Regression Results

Variable	Description
Intercept	Intercept
AGE	Age
PRODUCE	When I buy salty snacks, I also think of buying beer/wine.
BOGOSAVE	I don't believe that "2 for 1" salty snack deals save you much money.
BUGLES	Bugles is an acceptable brand to me.
STOCKUP	When I buy more salty snacks, I stockpile them.
HARDJUDG	The differences between salty snack brands are hard to judge.
…	…

Conclusion, Limitations, and Future Research

In this study, I developed a machine learning approach comprising a MHMM clustering model and a Lasso regression model to identify dynamic segments based on sequences of multiple item purchases. I applied the approach to data from the salty snacks category. I summarized the transaction data in a way that enabled MHMM to identify clusters of purchase sequences. The results show good descriptive power for the method used. In the predictive part of my approach, I used Lasso regression to predict cluster membership probabilities based on more than 200 independent variables in my dataset. The results suggest that the Lasso regression model has good predictive power. By understanding the segments and using the coefficients in the Lasso regression, managers can target the right customers with the right multiple item combinations.

My analysis has limitations that future research can address and extend in different ways. For instance, in the predictive part of the model, I focused on predictive accuracy. However, it is possible to use techniques like cluster-wise regression to obtain

descriptive profiles of customer segments. A challenge to this alternative technique is the large number of attitudinal and demographic variables. The literature on clustering high-dimensional data can provide guidance and it is possible to adopt some of the algorithms to segment customers based on both independent and dependent variables. Another possible way to extend this research is to develop a recommender system based on the proposed model. Such a system would recommend the next set of items to purchase for each consumer. My goal in this essay was to identify clusters of customers and predict future purchases, but it is also possible to modify this model to predict the next purchase of customers without assigning them to any one cluster. This extension can also be a powerful validation for my research.

CHAPTER IV

TOPIC HIDDEN MARKOV MODEL (THMM): A NEW MACHINE LEARNING APPROACH TO MAKE DYNAMIC PURCHASE PREDICTIONS

Introduction

To maximize the lifetime values of their customers, retailers and manufacturers increasingly desire accurate product recommender systems. Marketers from these firms strive to recommend the right products to the right customers for the right purchase occasion. Correct recommendations also help lower excess inventory, prevent stockouts, and reduce supply chain costs. Marketers seek recommender systems that can personalize recommended products to the customers with a high degree of accuracy.

The proliferation of customer data is enabling greater personalization in recommendations. Marketers are awash with data that include data on customer purchases and customer behavior on different media. To leverage these data and personalize product recommendations to customers, marketers are turning to newer personalization and recommender systems. The success of these systems depends on how accurately they predict customers' purchase behavior.

Prediction is thus key to personalization. However, this prediction task is extremely challenging for a retailer as it typically offers several hundreds or thousands of products, and a recommender system needs to recommend a short list of items to each customer on each purchase occasion based on accurate predictions of what the customer will likely purchase. Given the enormity of this challenge, retail managers do not sufficiently leverage customer purchase data to better predict their next purchases.

Traditional models that predict purchases have important limitations. Conventional models such as multinomial logit models of brand choice (e.g., Guadagni Little 1983) or purchase quantity models (e.g., Gupta 1988) rely on a few product characteristics and marketing variables that are often impractical to implement at a customer level. They are also estimated on small samples and often do not scale well. In reality, most retail firms have data on millions of customers and desire personalized recommendations, so analysis of a sample of customers is not that useful.

With data proliferation and advances in computing power and speed, machine learning models have captured the attention of marketers. Recommender systems that are not model based but based on co-occurrences of purchases of different items (e.g., collaborative filtering [CF]) are widely used in practice. Of late, topic models such as probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA) model have been applied to predict purchases (Blei et al. 2003; Hofmann 2001; Jacobs et al. 2016).

However, even these newer models have limitations. Collaborative filtering is count-based and has data sparsity issues (Su and Khoshgoftaar 2009). PLSA (Hofmann 2001), LDA (Blei et al. 2003), and their applications (e.g., Jacobs et al. 2016) are primarily static and combine all transactions for a customer. By incorporating purchase dynamics, prediction accuracies can be improved substantially. Purchase dynamics can be modeled through hidden Markov Models (Netzer et al. 2008; Zhang et al. 2014). However, HMM has not been used in the purchase prediction or recommendation context.

I creatively combine topic modeling and dynamic purchase modeling to develop a new machine learning model called the Topic Hidden Markov Model (THMM) to better predict, target, and personalize product recommendations. The proposed model captures purchase dynamics and customer heterogeneity for personalized prediction. It is superior to alternative models in two ways. It can capture dynamic or time-varying patterns in customers' purchase habits. It can learn patterns in transactions of multiple items in a purchase occasion (common in categories such as salty snacks and cereals).

My model significantly extends existing purchase prediction models. The fundamental data generation process in the proposed model is similar to those of PLSA or LDA. I introduce dynamics through a hidden Markov process on the topics or preferences/motivations that are effective in each shopping occasion. This results in common topics that are truly representatives of the general behavior of customers. I also allow for customer heterogeneity by assigning an idiosyncratic (or individual-specific) topic to every customer. The advantage of the idiosyncratic topics is that they let us capture the unique taste of each customer. Therefore, we obtain more accurate predictions when we use the common topics and idiosyncratic topic to predict customers' behavior.

I train my model as well as the three benchmark models, PSLA, LDA, and CF using a large dataset of customer transactions in the salty snacks category from a large retail chain. I predict the next purchases of each customer on a holdout validation sample and demonstrate the superior performance of the model over the benchmark models.

Related Literature

The topic of recommender systems has been investigated in multiple disciplines which have focused on different aspects of these systems (for reviews, see Adamavicius and Tuzhilin 2005; Breese et al.1998; Brusilovski et al. 2007; Jannach et al. 2011). I focus on the algorithms that predict customers' purchases or generate product recommendations[4].

Modeling customer choice or purchase has been of paramount importance to marketing researchers and practitioners. However, traditional marketing choice models (Guadagni and Little 1983; Gupta 1988; McFadden 1986), cannot be effectively used in typical recommender systems for a variety of reasons. First, traditional choice models do not scale up to the size of such applications (Naik et al. 2008). Second, recommending new items (items that have not been purchased or seen by the customer) is not straightforward using traditional models. Third, these models usually need some data on product, customer characteristics, and marketing mix variables, which may not be always available. Fourth, these models use a sample of customers for estimation, while the prediction task is intended for each customer of the firm.

Computer scientists, in particular, machine learning researchers, have approached the recommendation problem from the design of algorithms that efficiently and accurately generate recommendations. Content-based filtering methods are among the

---

[4] Other aspects of recommender systems studied include: (1) Aggregate-level as it is relates to the effects of recommender systems on firms, market, and society (Brynjolfsson et al. 2011; Fleder and Hosanagar 2009); and (2) Individual-level behavioral effects (Cooke et al. 2002; Senecal and Nantel 2004).

first class of algorithms proposed (Lops et al. 2011; Mooney and Roy 2000). These models recommend items with characteristics similar to those of items that a customer has purchased in the past. Content-based models rely heavily on product characteristics and attributes, which make them less applicable in retail settings with large assortments of products. Moreover, recommendations generated by these algorithms are generally limited to items that are similar to the items that a customer has already purchased, rendering recommendation of infrequently purchased or new items difficult.

Collaborative Filtering (CF) is another approach that is widely used in practice for generating recommendations (Adamavicius and Tuzhilin 2005; Breese et al. 1998; Brusilovski et al. 2007; Jannach et al. 2011). Modern recommendation methods are largely based on the CF perspective. This approach comprises analysis of past behavior or the opinions of existing customers for predicting which products the current (active) customer will be interested in purchasing (Jannach et al. 2011). In this approach, the main assumption is that users will continue to adopt their past behavior in the future.

The first group of collaborative filtering algorithms were instance-based techniques (Adamavicius and Tuzhilin 2005). This approach creates a matrix of customers' ratings of available products and use it to find similar customers and/or products. A commonly used instance-based CF technique is called user-based nearest neighbor recommendation (Jannach et al. 2011). In this method, given the purchase history of the current (active) customer, the goal is to identify other customers who have preferences similar to those of the active customer. It makes predictions about the active customer's preferences toward products she may not have seen yet based on the past

purchases of peer customers (Jannach et al. 2011). A big advantage of CF techniques is their simplicity and unlike content-based models, they require data only on past preferences and do not rely on demographic or product characteristics data. However, CF techniques suffer from issues related to data sparsity especially when there is a large assortment of products (Jannach et al. 2011).

An interesting example of a user-based nearest neighbor method is the model developed by Lu et al. (2016). They utilize video data of customers trying new garments to infer their preferences based on their facial expressions. In the next step, they use a user-based CF technique to recommend items based on the shopping behavior of like-minded customers.

Bodapati (2008) argues that all the work in the recommender systems literature calculate purchase probabilities conditional only on customer purchase history and ignore the impact of the recommended actions. Therefore, he develops a model that explicitly takes recommendations into account while calculating conditional purchase probabilities. Although his model outperforms benchmark models, data on previous recommendations are not always available. Even if they are available, they could be extremely sparse in large applications as recommender systems do not recommend many products to customers. The model also relies on product attributes data on which may not be always available.

Ansari et al. (2000) develop a content-based recommender system that needs different types of data, including customers' expressed preferences and demographics,

expert recommendations, and product characteristics to recommend items to customers. The data requirement is limiting in large settings.

Ansari et al. (2018) develop a probabilistic model that relies on user-generated content to predict users' ratings of movies. Specifically, they extract topical themes from user-generated tags about movies and use those topics as predictors of viewer ratings. The data requirements of this model are quite high as they require product covariates, user-generated textual descriptions, and user ratings as inputs. Furthermore, in the movie context, user preferences may change over time and their model cannot capture those dynamic patterns.

Fleder and Hosanagar (2009) develop a novel analytical model that mimics a simple recommender system based on a collaborative filtering design. The results of their model show that such recommender systems could potentially lead to lower sales diversity at an aggregate level as they tend to recommend more popular items to everyone.

A class of probabilistic models developed for purchase prediction are inspired by topic modeling techniques (Steyvers and Griffiths 2007). They use the analogy of modeling topics to model customer purchases. The idea in topic modeling and its application in purchase prediction is that, given a collection of textual documents (customers' purchase histories), each document exhibits a mixture of topics, where a topic is a probability distribution over a finite set of words (products). The biggest advantage in this approach is that topics provide a low-dimensional representation of the

data that allows the discovery of semantic relationships that could result in better prediction accuracy when data exhibit sparsity.

Another group of collaborative filtering models are called probabilistic models or model-based recommender systems. In this stream of research, researchers use tools and techniques from probability theory for modeling customers' preference data to address the purchase prediction problem (Hofmann 2003; Jacobs et al. 2016). In this approach, customers' preferences are modeled as stochastic events. The probability distribution of these events is analyzed to make inferences or predictions about customers' purchase preferences. Most of the model-based approaches deal with the data sparsity problem by allowing a mixture membership mechanism (Hofmann 2003; Jacobs et al. 2016). The idea is that if we find some common purchase (or preference) patterns, we could associate every customer behavior with each of those general patterns with a degree of certainty. This is usually done by introducing a finite set of hidden variables that capture general patterns.

For example, Jacobs et al. (2016) develop a purchase prediction model based on Latent Dirichlet Allocation (LDA) (Blei et al. 2003). They report a higher prediction accuracy than benchmark models that include collaborative filtering, mixture of Dirichlet distributions, and logistic regression. Another purchase prediction model inspired by topic modeling is Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 2001). These methods have been found to be effective in predicting customers' next purchases and to be efficient in working with large assortments of products. The findings

from these models are promising, calling for further research on recommender systems inspired by topic modeling techniques.

However, the topic models assume that a customer's preference does not change over time. This static assumption may not be always true, especially in product categories consumed repeatedly over time. Customer preferences often evolve with the age of the customer or with changes in the environment and the market (Sahoo et al. 2012). This is especially important when we are making predictions for product categories like salty snacks, cereals, and carbonated soft drinks from which customers usually buy more than one item in a transaction or shopping trip. There are complementary patterns among products in a shopping basket that cannot be captured using static models as these models ignore the time of purchase and pool a customer's transactions. Therefore, allowing for dynamics can improve the predictive power of the model.

In the recommender systems literature, a few papers incorporate temporal elements or capture dynamic patterns, but they do not follow a topic modeling approach. For instance, Rendle et al. (2010) propose a set of customer-specific Markov chains to model customer choices in different time periods. Their approach, however, suffers from extreme data sparsity issues as they need to estimate a transition matrix per customer. Sahoo et al. (2012) propose a model based on a hidden Markov model and a negative binomial mixture of multinomial distributions as the observation mechanism. Although their model outperforms other static benchmark models, its observation mechanism fails to account for heterogeneity in customers preferences.

I fill this gap by developing a purchase prediction model based on topic modeling that explicitly captures how customer preferences change from one time period to the next in a realistic manner. My model allows for customer heterogeneity by adding individual-specific taste distributions. These features intuitively enhance prediction accuracy.

Models

The central idea behind my modeling approach is as follows. Each customer's purchases of items within a transaction are correlated. So are her purchases over time. So are purchases across customers. I assume each customer has a hidden preference class or motivation (e.g., flavor-conscious, health-conscious, environment-conscious) for every transaction based on consumer preference theory. I begin by applying topic modeling to the problem treating customer purchase history as a document. In the language of topic modeling, products/items are words, customers are documents, and topics are preference classes/motivation segments. I overlay a hidden Markov model that incorporates purchase dynamics. I assume customer purchases reflect a mixture of preference classes. I exploit the fact that number of items purchased is typically far less than the number of words in a document to gain prediction accuracy, while incurring negligible additional computation cost. Unlike prior topic models that combine the transactions for a single customer over time, my THMM analyzes every transaction, capturing even the correlations among items bought together.

I now describe my dynamic model and compare it with benchmark models. I start by introducing the PLSA and LDA models that serve as comparators for my

dynamic model. I explain the different modeling assumptions as well as the likelihood estimation techniques.

I use the following notation. There are j = 1,…, J products in a store. For each customer i = 1,…, I, we observe a history of transactions over t = 1,…, T time periods. Every transaction of customer i is a set of products/items with C_{it} items in the transaction at time t. There could be multiple quantities of one product/item in a transaction and each of them contributes to the total number of items in the transaction. Also, let $N_i = \sum_{t=1}^{T} C_{it}$ be the total number of items purchased by customer i during her entire purchase history.

PLSA and LDA

Probabilistic Latent Semantic Analysis (PLSA) is a topic modeling technique (Hofmann 2001). The application of PLSA in purchase prediction is motivated by two reasons. First, the PLSA is a technique for the analysis of co-occurrence data over a discrete dyadic domain. It maps high-dimensional count vectors to a low dimensional representation in a latent space. Although, it has been mostly used in text analysis, the method could be used in purchase prediction application as purchase logs of customers could be transformed to count vectors. Second, in the text analysis application of PLSA, the high-dimensional count vectors of words in a document mapped to a lower dimensional semantical or topical level. This semantical level represents the intention of uttering a word (Hofmann 2001). In the context of purchase prediction, one could argue that customers' choices of products are driven by latent motivations or preferences (Jacobs et al. 2016). In this analogy, products, customers purchase histories, and

motivations in purchase prediction context are equivalents of words, documents, and topics in the text analysis context. For the remainder of this chapter, I use the terms motivation and preference class interchangeably to refer to the latent low-dimensional space.

Assume there are k=1,…,K preference classes or motivations. Each motivation is represented by a multinomial distribution φ_k over the available number of products J. Therefore, if a product purchase is motivated by motivation k, the probability of buying product j is equal to the jth element of vector φ_k, i.e., φ_{kj}.

A key assumption in PLSA is that each product purchase is driven by a single motivation. However, the entire purchase history of a customer may be driven by more than one motivation. For customer i, multinomial distribution θ_i over K motivations determines the relative frequency with which different motivations drive the purchases. For example, the probability that a product purchase of customer i is motivated by motivation k is equal to the kth element in θ_i, i.e., θ_{ik}.

I now introduce the data generation process of PLSA that helps summarize the assumptions about the model and the relationships between the variables. Imagine we want to generate the entire purchase history data. The generation process can be summarized as follows.

For each customer i, for each product purchase

Choose a motivation z ~ Multinomial (θ_i)

Choose a product w ~ Multinomial (φ_z)

Figure 4 shows the graphical model associated with the above data generation process. Note that w is the only observed variable and hence depicted in gray color in Figure 4.

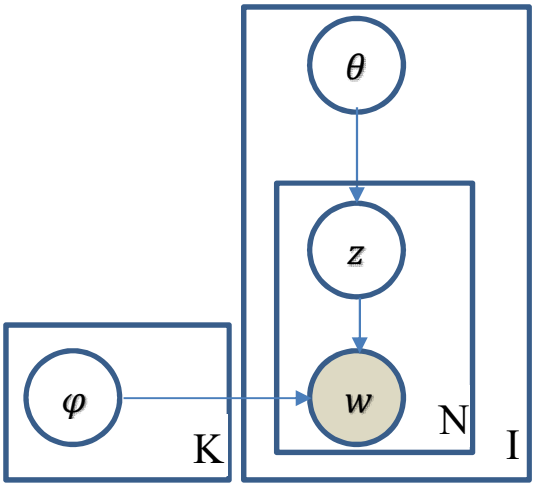

Figure 4 Graphical Model of PLSA

Therefore, we could write the likelihood of observing the whole data set D:

$$P(D|\Phi,\Theta) = \prod_{i=1}^{I}\prod_{w=1}^{N_i}\sum_{k=1}^{K}\theta_{ik}\times\varphi_{kw}$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\sum_{k=1}^{K}\theta_{ik}\times\varphi_{kj}\right)^{n(i,j)} \qquad (10)$$

Where $n(i,j)$ is the number of times product j has been purchased by customer i.

I use the direct numeric maximization or Expectation Maximization (EM) algorithm (Dempster et al. 1977) to estimate the parameters of this model. I employ the EM algorithm that starts with a set of initial values of parameters and alternates between

calculating the expected values of latent variables (z) and finding parameters that maximize the conditional likelihood, repeating the process until convergence.

Latent Dirichlet Allocation (LDA) is a topic modeling technique introduced by Blei et al. (2003). The data generation assumption of this model is similar to that of PLSA. The difference is that, in LDA, the parameters, θ_i and φ_k are assumed to be draws from Dirichlet distributions with parameters α and β, respectively:

Choose $\varphi_z \sim Dir(\beta)$

For each customer i, choose $\theta_i \sim Dir(\alpha)$. For each product purchase, choose a motivation z ~ Multinomial (θ_i) and choose a product w ~ Multinomial(φ_z), where $Dir(\alpha)$ is a Dirichlet distribution with parameter α. The inference for this model involves the estimation of parameters α and β. The most common estimation technique used in the literature is based on a collapsed Gibbs sampling algorithm (Griffith and Steyvers 2004).

THMM

I now introduce my dynamic model by starting from the PLSA model assumptions. The main objective is to include dynamics in the PLSA model. To capture dynamics, we should treat every transaction of a customer separately. Remember that each transaction can include multiple items. Moreover, recall that we have K motivations that capture the general purchase behavior of the entire data set. Here, I call them general or common motivations. I assume that all product choices in a transaction are driven by a single general motivation. The effective general motivation could differ

for the other transactions of a customer. I assume that these general motivations follow a Markov process over customers' shopping occasions.

Using a Markovian process to model customer choice is common in the marketing literature (Netzer et al. 2008). A theoretical justification for using Markov processes in modeling customer choice over time is based on "state dependence" (Dube et al. 2010). State dependence implies that current customer behavior is related to previous customer behavior. To illustrate this property, assume that there are two choices available to a customer, Product A and Product B. Positive state dependence suggests that the customer gains more utility from choosing Product A, if she had chosen Product A in the previous purchase occasion. Another type of state dependence is variety seeking, which implies that Product B would offer more utility to the customer if her previous choice was Product A.

I also introduce a set of individual-specific or idiosyncratic motivations to the model. I assume that for every customer i, there is a multinomial distribution η_i over products that captures the customer's unique taste or preferences. This assumption serves two purposes. First, it allows for customer heterogeneity in the model. Second, the resulting general motivations would be free of outliers, and would therefore be more generalizable to the entire population. Consequently, including these individual-specific motivations in the model can lead to a higher prediction accuracy.

We need a mechanism to determine whether the individual-specific motivation or one of the common motivations is driving a purchase. If a common motivation is driving a purchase, we need to determine which one of the K motivations is doing that. I assign a

binomial distribution to each customer with a parameter δ_i that reflects the probability that a customer's product choice is driven by a general motivation rather than the individual-specific motivation. This parameter also indicates the degree to which a customer deviates from general motivations.

I assume that general motivations follow a Markov process. Since the motivations are hidden, I could also refer to the model as a variant of a hidden Markov model (HMM). Let q_t be the general motivation that drives customer i's purchases at time t[5]. The Markov process of general motivations has the following properties:

$P(q_s|q_{s-1}, \dots, q_1) = P(q_s|q_{s-1})$, i.e., the current general motivation depends only on the previous general motivation.

$a_{uv} = P(q_t = v|q_{t-1} = u) \; \forall \; t = 2, \dots, T$ shows the transition probability from general motivation u to general motivation v. These transition probabilities are elements of matrix A.

$\pi_u = P(q_1 = u)$ is the probability that the initial motivation (motivation in the first transaction) is u. Vector π contains all these probabilities.

Therefore, I could describe the model using the following data generation process.

For each transaction of customer i, choose a general motivation q from the hidden Markov model with parameters (A, π) based on the previous transaction's

[5] For expositional ease, I drop the customer index from this and the subsequent expressions.

general motivation. For each product purchase in the transaction, choose between general or individual-specific motivations based on z ~ Binomial(δ). Choose a product w ~ Multinomial (φ_q) or Multinomial (η). If C_t is the number of items in the t-th transaction and w_s is the product that have been chosen in the sth product choice in a transaction, the probability of observing a sequence of purchases for a customer is equal to:

$$P(D|\Phi,\Theta) = \sum_{all\ q} \pi_{q_1} \prod_{s=1}^{C_1} (\delta\phi_{q_1,w_s} + (1$$

$$- \delta)\eta_{w_s}) \prod_{t=2}^{T} (a_{q_{t-1},q_t} \prod_{s=1}^{C_t} (\delta\phi_{q_t,w_s} + (1-\delta)\eta_{w_s})) \quad (11)$$

Note that q is a hidden variable and the summation in the above likelihood function shows that the probability should be calculated for every possible sequence of the K general motivations over T time periods. A graphical depiction of the proposed model appears in Figure 5.

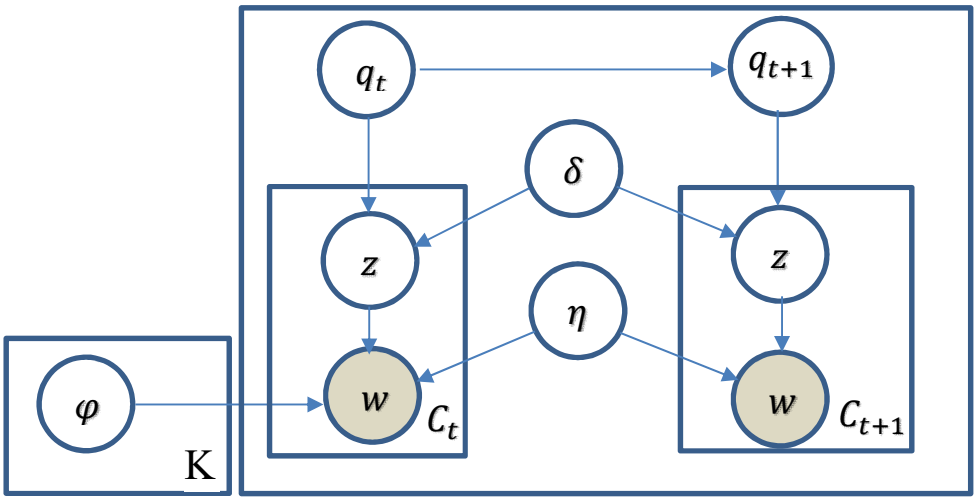

Figure 5 Graphical Model of THMM

A comparison of Figure 4 and Figure 5 shows how the proposed dynamic model captures the relationship between transactions at different time periods using a hidden Markov model. Note that in PLSA, all transactions are pooled together and there is no distinction among purchases at different time periods. Such treatment of data also ignores any information that might exist in relationships between items purchased in one transaction.

I estimate the parameters of the dynamic model using maximum likelihood estimation. Because there are latent variables in the model, I apply an Expectation Maximization technique called the Baum-Welch algorithm (Zucchini et al. 2016). The Baum-Wlch algorithm is an EM variant traditionally used to estimate hidden Markov models.

Collaborative Filtering

As a benchmark, I develop a user-based collaborative filtering model. User-based collaborative filtering has been extensively used in research as well as in many commercial recommender systems (Konstan et al. 1997; Resnick et al. 1994). In this approach, each customer belongs to a group of customers that behave in a similar manner. This assumption implies that products frequently purchased by other members of the group could serve as potential recommendations to a focal member.

I use a version of user-based CF called top-N recommendations. In this method, we have an active customer for whom we want to compute the top-N recommendations. Assume that R is an $(I-1) \times J$ user-item matrix containing purchase information of every customer but the active customer. In this matrix, element r_{ij} is one if the ith

customer has purchased the jth item in the past, and zero otherwise. User-based CF methods work in the following way to recommend the top-N items to the active customer.

The first step is to identify the v most similar customers to the focal customer in the data. To this end, I treat every customer (including the active customer) as a vector in the J-dimensional product space (Sarwar et al. 2000). I measure the similarity between the active customer and other customers by computing the Jaccard similarity coefficient between these vectors. After finding the v most similar customers based on the similarity measure, I aggregate their corresponding rows in R to find the set x of items purchased by the group. I then could find the N most frequent products in x to recommend to the active customer.

Data

I obtain data on the salty snacks category from a large New York state retailer that has loyalty card data on customer purchases but makes store level marketing decisions based on aggregate data. The retailer seeks a dynamic personalized marketing recommendation system with a high degree of accuracy. A leading salty snacks brand manufacturer, a Fortune 500 consumer packaged goods company, also makes aggregate targeting decisions for its salty snacks products but seeks a personalized model to market the right products to the right customers in the right week.

I use a panel data set of households' transactions in the salty snacks category to estimate the proposed model and compare it against the benchmarks. My data include

purchase histories of 2,888 households over a one year period. Table 7 provides a summary of the data.

Table 7 Summary of Data

Variable	*Number of Observations*
Number of UPCs	480
Number of transactions	73,086
Number of customers	2,888
Average number of transactions	25.31
Average number of items in a transaction	1.91
Percentage of multiple item transactions	56%

Two numbers in Table 7 are worth noting. First, there are 480 different universal product codes (UPCs) in the data set. I use the raw UPCs from the data set without any grouping based on brands (unlike Jacobs et al. (2016) who assign the same UPC to different sizes of the same brand). I improve on Jacobs et al.'s (2016) grouping that may not be accurate as customers differentiate among different sizes of the same brand. Second, about 56 percent of all the transactions involve purchases of multiple items in a category. This finding shows that the capability of a model to analyze multiple transactions could be crucial.

I split the customers into two groups: one for model selection and the other for training and testing. Table 8 shows this division. For each group, I drop the last transaction of every customer and use the remaining transactions to train my model. I first analyze the model selection data to determine the optimal number of general

motivations (K) in the models. After calibrating the models, I do the main performance comparisons using the train and test data.

Table 8 Sample Sizes

Sample Type	*Sample Size*
Validation	888
Train and test	2,000
Total	2,888

Results

To tune and compare my models, I need a prediction accuracy measure. I use the weighted hit rate, consistent with Jacobs et al. (2016). I calculate the weighted hit rate by looking at a set of products, called a prediction set. Based on the model predictions, these products have the highest chances of being purchased by a customer. I compare the products in a prediction set against the actual transaction made by the customer. The goal is to achieve the highest number of matches with a very small prediction set size. Because the positions of products in the prediction set are important (Xu and Kim 2008), I incorporate a weighting mechanism consistent with Jacobs et al. (2016).

I define the weighted hit rate as follows. A prediction set r_i of size S for customer i contains the ordered list of the S highest ranked products for that customer. Assume that y_i is the last transaction of customer i we are going to predict. The number of unique items in this transaction is u_i. Moreover, r_{i1} is the first element in the prediction set, i.e., the product with the highest purchase probability for the model-based rankings and the

highest product score for the collaborative model. I also use the function $w(s,S) = 1 - (s-1)/S$ to weight a hit for the s-th ranked product in a prediction set of size S. Thus, the weighted hit rate is calculated as follows:

$$h_i(r_i,S) = \frac{\sum_{s=1}^{S} I[r_{is} \in y_i] w(s,S)}{\sum_{s=1}^{\min(S,u_i)} w(s,S)} \qquad (12)$$

The number of general motivations in PLSA, LDA, and THMM is an input to the algorithms and should be determined prior to training the models. Recall that we need to determine the number of motivations, K, before training the models. To this end, I adopt the following process. I drop the last transaction of every customer in the model selection sample and train the models using the remaining customers' transactions. I performed the training for different values of K, starting from K=2. For each value of K, I calculate the average hit rate based on prediction set sizes 1 to 25. For each model, I pick the value of K that results in the first local maximum in average hit rate. In other words, if avhr(K) shows the average hit rate for K motivations, the optimal K or K* is the lowest number for which avhr(K*-1)< avhr(K*)< avhr(K*+1). This process is consistent with the procedure of Jacobs et al. (2016) to determine the number of motivations, with the key difference being that they use the average predictive likelihood of their models as the objective function. My results show that the optimal number of motivations for THMM, PLSA, and LDA is the same and is equal to 5. The average weighted hit rates for different values of K appears in Figure 6.

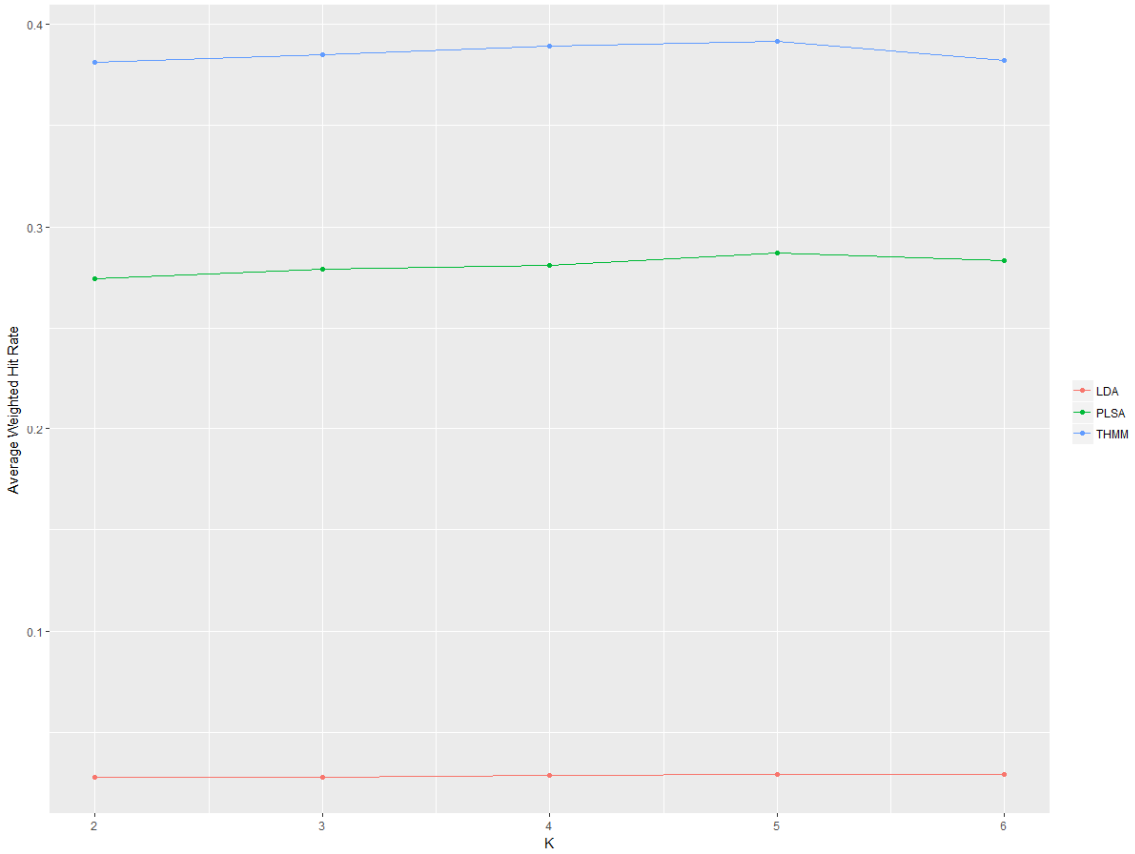

Figure 6 Weighted Hit Rate as a Function of Number of Common Motivations

Figure 7 shows the average Weighted hit rate for the models. Note that for each prediction set size, the hit rate in Figure 7 is an average across all the customers in the sample. Figure 7 shows the superior performance of THMM over PLSA, LDA, and CF.

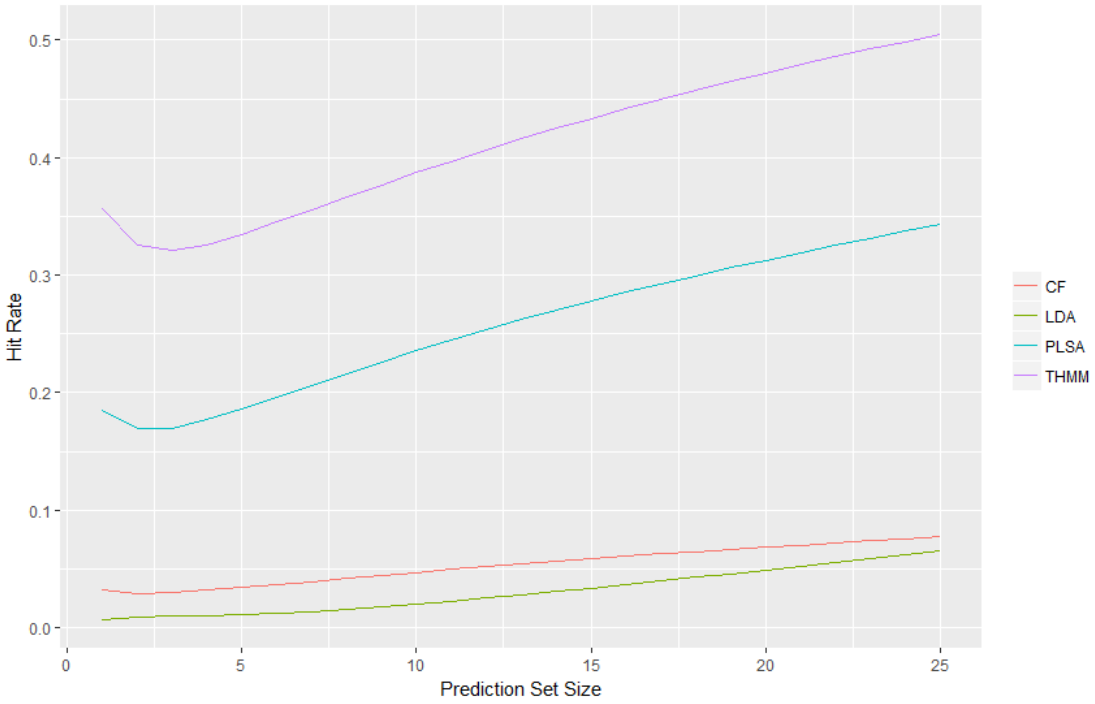

Figure 7 Hit Rate Comparison of THMM and Benchmark Models

To test the robustness of the models, I estimated all the models with various sub-samples of the data. The accuracy of the model in a different data scenarios appears in Figure 8. Figure 8 shows the hit rates when we observe a short purchase history for every customer. To perform this analysis, I limited my sample to customers who made 12 to 20 transactions in the period of investigation. This means that we have a limited number of transactions to learn the behavior of customers. We could see that THMM's predictive performance is still superior to those of the benchmark models. However, the performance gap between THMM and PLSA decreases with full data (Figure 7). This finding suggests that THMM needs more training data to achieve its peak performance, which makes sense because THMM has more parameters than PLSA.

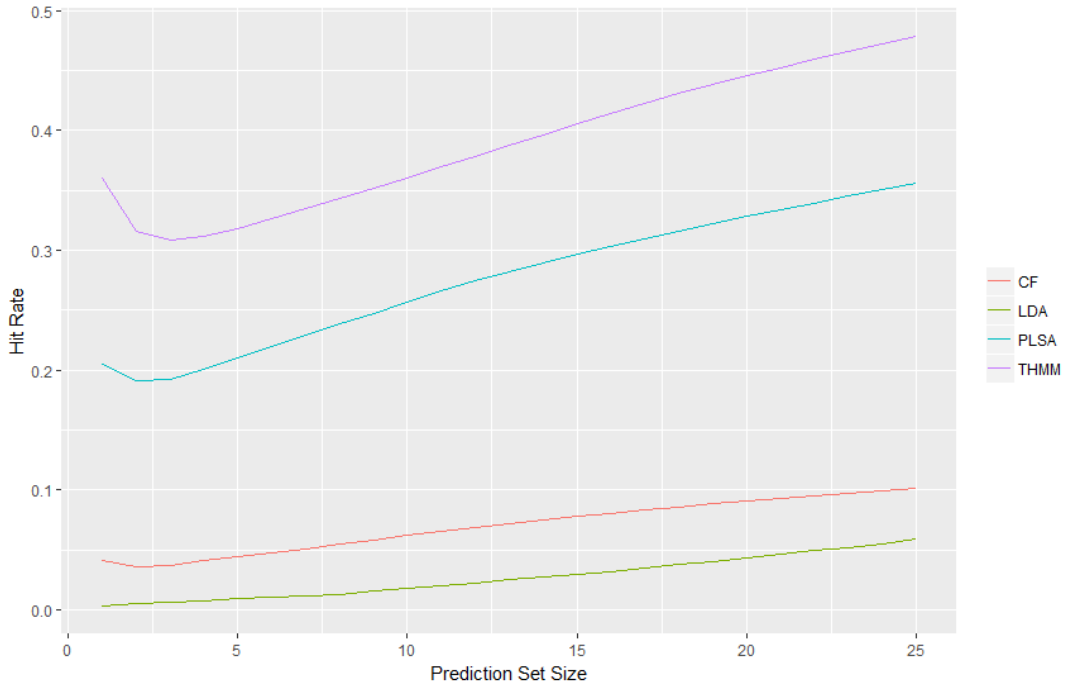

Figure 8 Hit Rate Comparison of THMM and Benchmark Models (for Customers with Few Transactions [12-20])

The next data scenario I test is the opposite of the previous one and looks at the case when a longer purchase history is available for each customer. I expect to see an improved performance for THMM for the reason discussed in the previous paragraph. Figure 9 confirms my expectations as the gap between THMM and other models increases relative to Figure 7 and Figure 8. This finding suggests that THMM improves with a longer history of purchases.

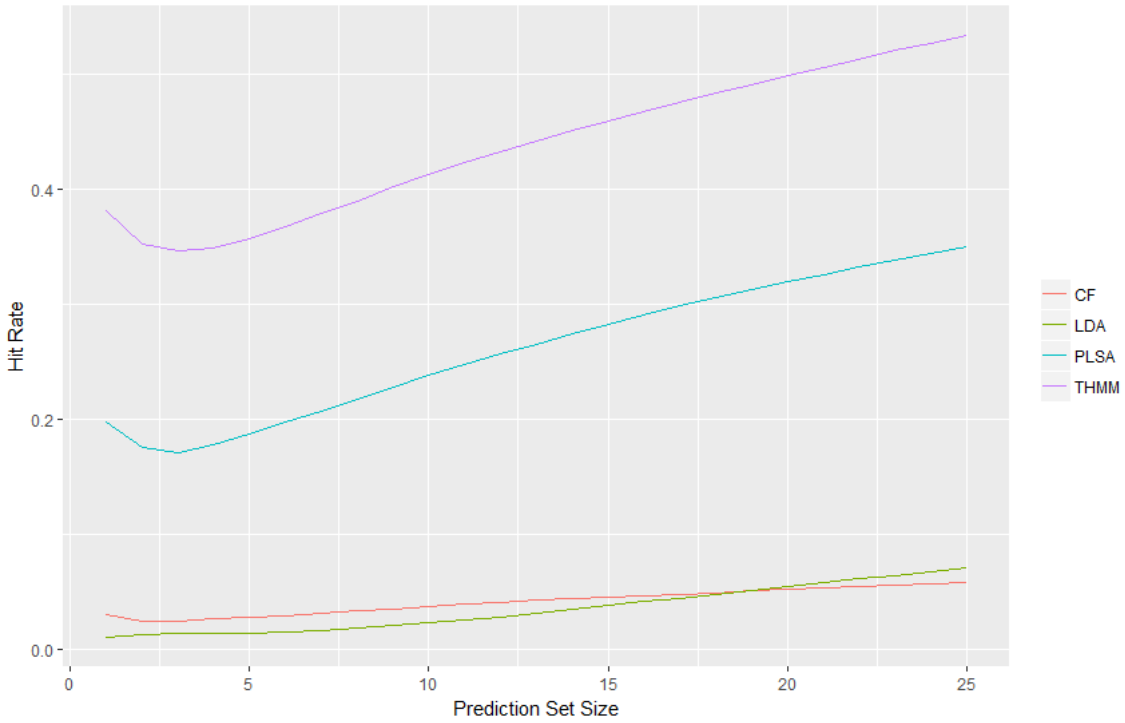

Figure 9 Hit Rate Comparison off THMM with Benchmark Models (for Customers with Many Transactions [21-80])

In the last scenario, I place some restrictions on the products rather than the customers. This is because there is a small fraction of products that accounts for a large portion of sales. While the prediction of the purchases of these products could be high for most models, it would be interesting to examine the accuracy of the models in predicting the purchase of products that are not frequently purchased at the store. To do that, I drop 20% of the products in the assortment that are most frequently purchased by customers. Figure 10 shows the hit rate comparison for the models in this scenario. There is a significant performance gap between THMM and the benchmark models.

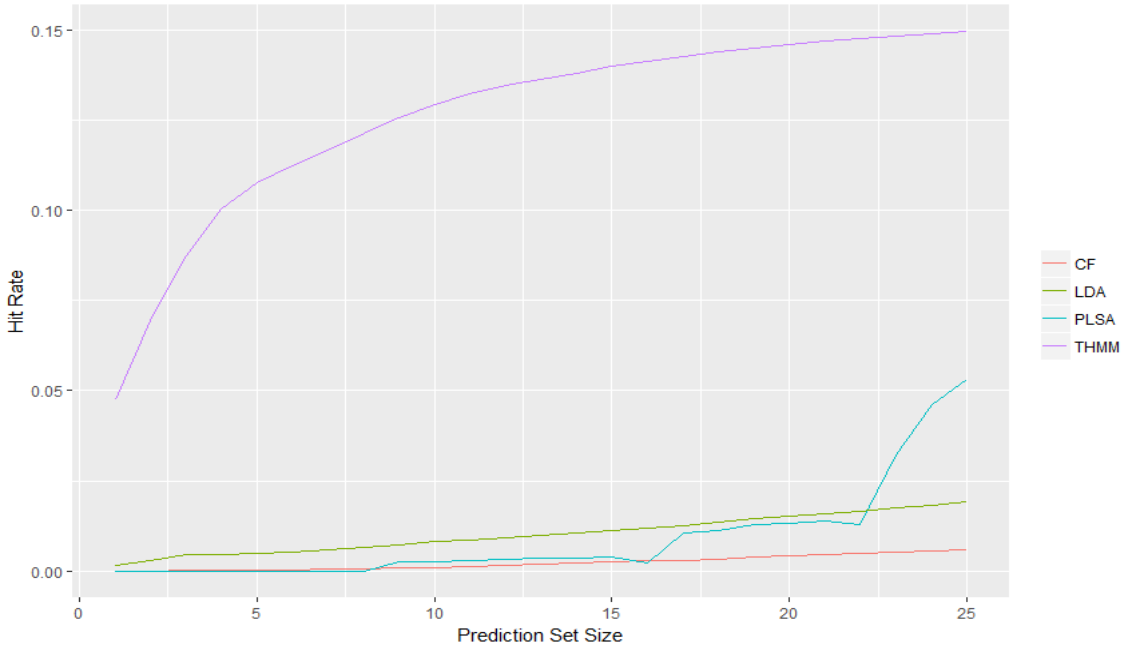

Figure 10 Hit Rate Comparison of THMM with Benchmark Models (for Infrequently Purchased Products)

THMM has the highest prediction accuracy (Weighted hit rate). THMM is superior in various robustness checks (different restrictions on the number of transactions and subset of products). The performance gap between THMM and PLSA decreases with as the number of customer transactions decreases. THMM needs more customer transactions to reach its peak performance. The opposite applies to collaborative filtering (CF). Customers with more transactions tend to show more variety-seeking behavior. Unfortunately, CF cannot find effective matches for these customers. The most difficult prediction task is predicting the purchases of infrequently purchased products. THMM is considerably more accurate than the benchmark models on this task.

Model Robustness Checks

To test that THMM is robust to related model specifications, I tested different variants of THMM by modifying my assumptions about δ and η of THMM. The results appear in the APPENDIX B. The results show that the proposed THMM offers the best performance and that the results are consistent across the different versions of THMM.

Implications for Theory and Practice OF Marketing

Implications for Theory

My approach and model have interesting implications for theory. First, my model is based on hidden or underlying motivations or preferences of customers. By uncovering these motivations, my model sheds light on the possible mechanisms leading to purchases. My method identifies revealed preferences that are important for researchers to further explore.

Second, by splitting the motivations into motivations common to a group and idiosyncratic motivations, my model offers deeper insights into what drives customer purchases. For example, a health-conscious motivation that is common to a group may lead customers in that segment to purchase low-fat, low-calorie items. An idiosyncratic motivation such as purchases of items with a red and green color packages could lead a customer to purchase any product that exhibits these colors. Understanding both these motivations is key to a complete analysis of customer purchase motivation.

My model shows that the incorporation of dynamics substantially improves prediction. The superior performance of my model relative to the benchmark models offers deeper insights into the role of dynamics in prediction accuracy. By including

dynamics, I am exploiting the variation in the purchases of items within a customer's transaction, across transactions, and across customers.

Finally, my model offers opportunities for improvement in prediction accuracy by incorporating covariates. Marketing mix variables and customer characteristics can be included in the model as covariates in the topic modeling component or the hidden Markov transitional probability or both. Such a theoretical advancement will lead to a complex likelihood function, but could be estimated using powerful computational methods. If data are available on these covariates, an enhanced version of THMM could be developed and estimated, possibly leading to higher prediction accuracy.

Implications for Practice

My model has huge implications for marketing practice. A typical supermarket carries about 40,000 stock keeping units (SKUs). A supermarket chain has a customer base ranging from a few hundred thousand to tens of millions of customers. Accurate purchase prediction and recommendation of products can substantially boost sales, lower supply chain and inventory costs, and enhance profitability. A one percent improvement in prediction accuracy for a retailer with one million customers each resulting in incremental sales of $4 per customer per purchase occasion can result in an annual sales gain of $2 million. Assuming inventory and supply chain costs to be 15% of sales revenues, this incremental revenue also translates into an incremental cost savings of $300,000 for every one percentage increase in prediction accuracy for this retailer.

The model offers retailers and manufacturers insights into the hidden preferences of customers. By identifying the hidden motivations, retailers and manufacturers can

create new products and market bundles of products. For example, if the common motivations of a group of customers is environment-consciousness, managers can recommend environment friendly products. They can also be recommended as a bundle. Marketers can also develop additional new products that are gentle for the environment and recommend them to customers belonging to this segment.

Both retailers and manufacturers can also leverage the learning from the identified idiosyncratic preferences of customers. For example, by learning that a customer makes purchases based on her preference for items whose packages display a specific color, a retailer can recommend more products (even if they are seemingly unrelated) whose packages also carry the same color. The discovery of many of these idiosyncratic preferences across customers will also serve as a rich source of new product ideas for marketers.

Conclusions, Limitations, and Future Research

In this essay, I proposed a new purchase prediction model called THMM that is inspired by topic modeling as well as dynamic modeling. This model has unique features that differentiate it from other purchase prediction techniques based on topic modeling. First, THMM captures the dynamic or time varying correlated purchase patterns in customer purchases. I achieve this by allowing customer motivations to change over shopping occasions based on a Markov process. Second, THMM is capable of finding patterns within a transaction. In some product categories, there are multiple instances of multiple item transactions with some complementarities between the items in a transaction. Third, THMM could capture the unique preference/motivation of every

customer as Well. This makes the general motivation free of outliers and thus more generalizable to other customers.

I tested my proposed model on a dataset of customers' transactions in the salty snacks category. I compared the prediction accuracy of THMM with that of PLSA, LDA, and CF models. My results indicate that THMM has a superior prediction accuracy in various prediction tasks, including cases with short and long purchase histories and those involving purchases of less popular items. I attribute this superior performance to the characteristics of THMM.

Thus, the new machine learning model builds on customer preference theory and combines computer science and statistics in a unique way. My method and results enable companies to better predict each customer's future purchase pattern, target the right customers for the right products, recommend the next products to purchase, and offer the right promotions for the right products to the right customers.

My model has limitations that future research could address. First, my model does not incorporate any covariates. Although the THMM's prediction accuracy is greater than those of benchmark models, THMM could be improved by incorporating covariates such as marketing mix variables.

Second, my empirical application focused on one product category. Developing a cross-category prediction model will be challenging with added complexity, but it would be worthwhile task for a retailer.

Third, I did not specifically address issues such as memory requirements and real-time updating of data because such issues fall under computer systems and outside

the scope of my research. However, these issues are important from a scalability standpoint. Future research could tackle these issues as well.

Finally, although my model uncovered hidden segments, my research did not focus on segmentation in particular. Segmentation of customers in sales response is important for marketers (Bucklin et al. 1998). Future studies could explore the hidden segments in greater detail, focusing on the formation and behavior of such segments.

CHAPTER V

CONCLUSION

In the two essays of this dissertation, I sought answers to two important marketing problems by using a machine learning approach. The first essay addressed the segmentation problem and the second essay looked at personalized product recommendation. The common theme between the two essays is that I used hidden Markov Model to be able to capture the dynamic shopping behavior of customers.

In the first essay, I proposed a segmentation framework based on a mixture of hidden Markov Models. My results show that there are segments of customers whose preferences do not change over time. But there are some dynamic segments as Well. For the dynamic segments, I showed that there is heterogeneity in the number of hidden states customers transition among. The Lasso regression formulation in the framework can effectively predict outcomes using a large set of demographic and attitudinal predictors. I show that this model has a high prediction accuracy. These results could open an avenue of research on sequence clustering and mixture dynamic models in marketing that account for heterogeneity in the number of states. Although I used data from salty snack category, this tool can be easily applied to other product categories.

In the second essay, I combined a topic modeling technique (PLSA) with a hidden Markov Model to build a recommender system that accounts for dynamic changes in customers preferences. Topic modeling techniques have been used in the prediction of customer purchases. However, the previous methods did not account for purchase dynamics and multiple item purchases together with customer heterogeneity.

My approach uses a dynamic purchase prediction method inspired by topic modeling. The results show that my new model, THMM, outperforms existing models in predicting product category sales by a wide margin. This model is an important tool in managers' arsenal for product recommendations that can lead to higher sales in the product category.

REFERENCES

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* (6): 734-49.

Ansari A, Essegaier S, Kohli R (2000) Internet recommendation systems. *Journal of Marketing Research* 37 (3): 363-375.

Ansari A, Li Y, Zhang JZ (2018) Probabilistic topic model for hybrid recommender systems: a stochastic variational Bayesian approach. *SSRN Working Paper*.

Bodapati AV (2008) Recommendation systems with purchase data. *Journal of Marketing Research* 45(1): 77-93.

Borah A, Tellis GJ (2016) Halo (spwellover) effects in social media: do product recalls of one brand hurt or help rival brands? *Journal of Marketing Research* 53(2): 143–160.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993-1022.

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 43-52.

Brusilovski P, Kobsa A, Nejdl W (2007) The adaptive web: methods and strategies of web personalization. Berlin: Springer-Verlog.

Brynjolfsson E, Hu Y, Simester D (2011) Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. *Management Science* 57(8): 1373-1386.

Bucklin RE, Gupta S, Siddarth S (1998) Determining segmentation in sales response across consumer purchase behaviors. *Journal of Marketing Research* May: 189-197.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research* 43(3): 345–354.

Chiu CY, Chen YF, Kuo IT, Ku HC (2009) An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications* 36(3): 4558–4565.

Chung J, Rao VR (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.

Cooke ADJ, Sujan H, Sujan M, Weitz BA (2002) Marketing the unfamiliar: the role of context and item-specific information in electronic agent recommendations. *Journal of Marketing Research* 39(4): 488-497.

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

Dempster AP, Laird NM, Rubi DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 39(1): 1-38.

DeSarbo WS, Manrai AK, Manrai LA (1993) Non-spatial tree models for the assessment of competitive market structure: An integrated review of the marketing and psychometric literature. *Handbooks in Operations Research and Management Science* 5: 193–257.

DeSarbo WS, Grewal R, Scott CJ (2008) A clusterwise bilinear multidimensional scaling methodology for simultaneous segmentation and positioning analyses. *Journal of Marketing Research* 45(3): 280–292.

Dew R, Ansari A (2018) Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science* 37(2): 216-235.

Dubé JP (2004) Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Science* 23(1): 66–81.

Dubé JP, Hitsch GJ, Rossi PE (2010) State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics* 41(3): 417-445.

Elrod T (1988) Choice map: Inferring a product-market map from panel data. *Marketing Science* 7(1): 21–40.

Elrod T (1991) Internal analysis of market structure: recent developments and future prospects. *Marketing Letters* 2(3): 253–266.

Elrod T, Keane MP (1995) A factor-analytic probit model for representing the market structure in panel data. *Journal of Marketing Research* 32(1): 1–16.

Erdem T (1996) A dynamic analysis of market structure based on panel data. *Marketing Science* 15(4): 359–378.

Farquhar PH, Rao VR (1976). A balance model for evaluating subsets of multiattributed items. *Management Science* 22(5): 528–539.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: the impact of recommender systems on sales diversity. *Management Science* 55(5): 697-712.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences,* 101 suppl 1, 5228-5235.

Grover R, Srinivasan V (1987) A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research* 24(2): 139–153.

Guadagni PM, Little J (1983) A logit model of brand choice calibrated on scanner data. *Marketing Science* 2(3): 203-208.

Gupta S (1988) Impact of sales promotions on when, what and how much to buy. *Journal of Marketing Research* 25(4): 342-355.

Han J, Fu Y (1999) Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering* 11(5): 798–805.

Harlam BA, Lodish LM (1995) Modeling consumers' choices of multiple items. *Journal of Marketing Research* 32(4): 404–418.

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics.

Henning-Thurau T, Wiertz C, Feldhaus F (2014) Does twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science* 43: 375–394.

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1-2): 177-196.

Hofmann T (2003) Collaborative filtering via Gaussian probabilistic latent semantic analysis. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *ACM*, 259-66.

Huang D, Luo L (2016) Consumer preference elicitation of complex products using fuzzy support vector machine active learning *Marketing Science* 35(3): 445-464.

Jacobs BJD, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. *Marketing Science* 35(3): 389-404.

Jannach D, Zanker M, Felfering A, Friedrich G (2011) Recommender systems: an introduction. New York, NY: Cambridge University Press.

Jaworska J, Sydow M (2008) Behavioral targeting in on-line advertising: an empirical study. *Web Information Systems Engineering* 62:76.

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

Kaufman L, Rousseeuw PJ (1987) Clustering by means of medoids. Y. Dodge (Ed.), Statistical data analysis based on the L1 Norm (North-Holland, Amsterdam) 405-416.

Kim J, Allenby GM, Rossi PE (2007) Product attributes and models of multiple discreteness. *Journal of Econometrics* 138(1): 208–230.

Kim SY, Jung TS, Suh EH, Hwang HS (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications* 31(1): 101–107.

Klastorin TD (1985) The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science* 31(1): 84–95.

Konstan JA, Mweller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J (1997) GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM* 40(3): 77-87.

Liu X, Singh PV, Srinivasan K (2016) A structured analysis of unstructured big data by leveraging cloud computing. *Marketing science* 35(3): 363-388.

Liu Y, Kiang M, Brusco M (2012) A unified framework for market segmentation and its applications. *Expert Systems with Applications* 39(11): 10292–10302.

Lops P, De Gemmis M, Semeraro G (2011) Content-based recommender systems: state of the art and trends. *Recommender Systems Handbook* Boston, MA: Springer, 73-105.

Lu S, Xiao L, Ding M (2016) A video-based automated recommender (VAR) system for garments. *Marketing Science* 35(3): 484-510.

MacDonald IL, Zucchini W (1997) Hidden Markov and other models for discrete-valued time series (Vol. 110) CRC Press.

Mahajan V, Jain AK (1978) An approach to normative segmentation. *Journal of Marketing Research* 15: 338–345.

McFadden D (1986) the choice theory approach to market research. *Marketing Science* 5(4): 275–297.

Mooney RJ, Roy L (2000) Content-based book recommending using learning for text categorization. *Proceedings of the Fifth ACM conference on Digital Libraries*, *ACM*, 195-204.

Mount DM (2004) Bioinformatics: sequence and genome analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Naik P, Wedel M, Bacon L, Bodapati A, Bradlow E, Kamakura W, Kreulen J, Lenk P, Madigan DM, Montgomery A (2008) Challenges and opportunities in high-dimensional choice data analyses. *Marketing Letters* 19(3-4): 201-213.

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Science* 27(2): 185-204.

Poulson, CS (1990) Mixed Markov and latent Markov modelling applied to brand choice behavior. *International Journal of Research in Marketing* 7: 5–19.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2): 257–286.

Rendle S, Freudenthaler C, Schmidt-Thieme L (2010) Factorizing personalized Markov chains for next-basket recommendation. *Proceedings of the 19th International Conference on World Wide Web*, *ACM*, 811-820.

Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) GroupLens: an open architecture for collaborative filtering of Netnews. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM, 175-186.

Rutz OJ, Trusov M, Bucklin RE (2011) Modeling indirect effects of paid search advertising: Which keywords lead to more future visits? *Marketing Science* 30(4): 646-65.

Rutz OJ, Sonnier GP, Trusov M (2017) A new method to aid copy testing of paid search text advertisements. *Journal of Marketing Research* 54(6): 885-900.

Sahoo N, Singh PM, Mukhopadhyay T (2012) A hidden Markov model for collaborative filtering. *MIS Quarterly* 36(4): 1329-1356.

Sarwar B, Karypis G, Konstan J, Riedl J (2000) Analysis of recommendation algorithms for E-commerce. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, *ACM*, 158-67.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* 36(4): 500-522.

Senecal S, Nantel J (2004) the influence of online product recommendations on consumers' online choices. *Journal of Retailing* 80(2): 159-169.

Smith WR (1956) Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing* 21(1): 3–8.

Späth H (1979) Algorithm 39 clusterwise linear regression. Computing 22(4): 367–373.
Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* (Methodological) 267-88.

Steyvers M, Griffiths T (2007) Probabilistic topic models. *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum, 427(7): 424-440.

Su, X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. *Advances in Artificial Intelligence* Article ID 421425, 19 pages.

Thieme RJ, Song M, Calantone RJ (2000) Artificial neural network decision support systems for new product development project selection. *Journal of Marketing Research* 37(4): 499-507.

Wedel M, Kamakura WA (2000) Market segmentation: conceptual and methodological foundations, Springer Science & Business Media.

Wedel M, Kistemaker C (1989) Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing* 6(1): 45–59.

Wedel M, Steenkamp JBE (1991) A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation. *Journal of Marketing Research* 28(4): 385–396.

Xu YC, Kim HW (2008) Order effect and vendor inspection in online comparison shopping. *Journal of Retailing* 84(4): 477-486.

Zhang J, Netzer O, Ansari A (2014) Dynamic targeted pricing in B2B relationships. *Marketing Science* 33(3): 317-337.

Zucchini W, MacDonald IL, Langrock R (2016) Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.

APPENDIX A

The Lasso (Least Absolute Shrinkage and Selection Operator) is a special regression analysis method. Although originally defined for least squares, Lasso regularization is easily applicable to a wide variety of statistical models including generalized linear models.

Consider a sample consisting of N cases, each of which consists of p covariates and a single outcome. Let y_i be the outcome and $x_i := (x_1, x_2, \dots, x_p)^T$ be the covariate vector for the ith case. Then the objective of Lasso is to solve

$$\min_{\beta_0,\beta}\{\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\} \text{ subject to } \sum_{j=1}^{p}||\beta_j|| \leq u.$$

Here u is a prespecified free parameter that determines the amount of regularization. Letting X be the covariate matrix so that $X_{ij} = (x_i)_j$ and x_i^T is the ith row of X, we can write this more compactly as:

$$\min_{\beta_0,\beta}\{\frac{1}{N}||y - \beta_0 - X\beta||_2^2\} \text{ subject to } ||\beta||_1 \leq u.$$

We can rewrite the above equation in the following form:

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{N}\|y - X\beta\|_2^2 + \lambda \| \beta \|_1\right\}$$

The tuning parameter λ controls the strength of the penalty.

APPENDIX B

In this appendix, I train alternative versions of THMM and test if the results are robust to these versions. I created these models based on reasonable modifications in assumptions about δ and η of THMM.

THMM2 (One Static Motivation)

In this model, I assume that instead of a series of individual-specific motivations, there is only one motivation shared by all customers. Note that in the original THMM model, each individual-specific motivation serves as a static motivation for a customer, i.e., a motivation that does not change over time. In THMM2, I assume that there is only one of these motivations and it is shared by all customers. The goal is to investigate whether allowing for heterogeneity improves the prediction accuracy of the model or it is an unnecessary assumption.

Figure 11 shows the graphical depiction of THMM2. Note how η is modeled differently in this new model.

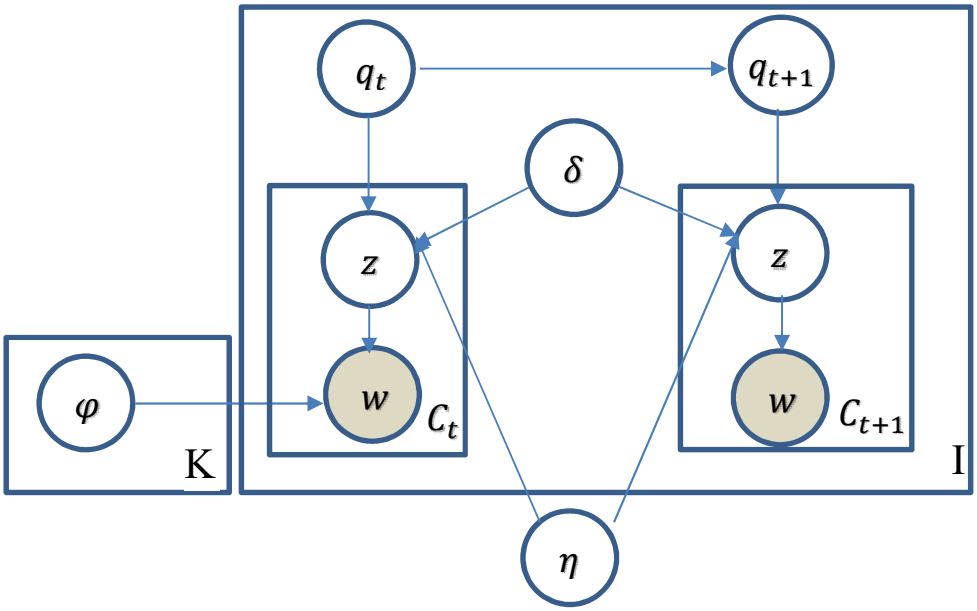

Figure 11 Graphical Model of THMM2

After training THMM2, I compare prediction accuracies based on the hit rate. Figure 12 shows the results. THMM has a significantly higher prediction accuracy than THMM2, suggesting that individual-specific motivations increase the prediction accuracy of the models.

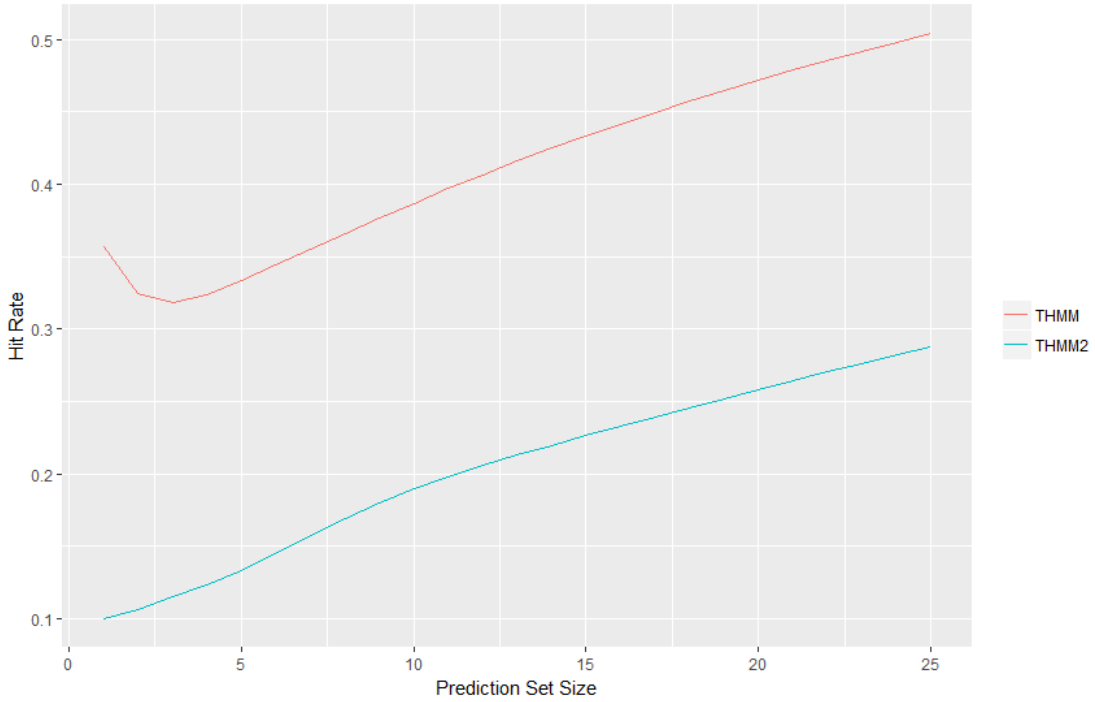

Figure 12 Hit Rate Comparison THMM and THMM2

THMM3 (Time-varying $\boldsymbol{\delta}$)

In this analysis, I allow the δ parameter to vary over time for every customer. This means that a customer's likelihood of deviating from the common behavior could change over time.

Figure 13 graphically depicts THMM3, and Figure 14 compares its hit rate with that of THMM. As it is evident from Figure 14, the performances of the models are very close to each other. However, THMM3 has more parameters, which causes estimation complexities and slows convergence.

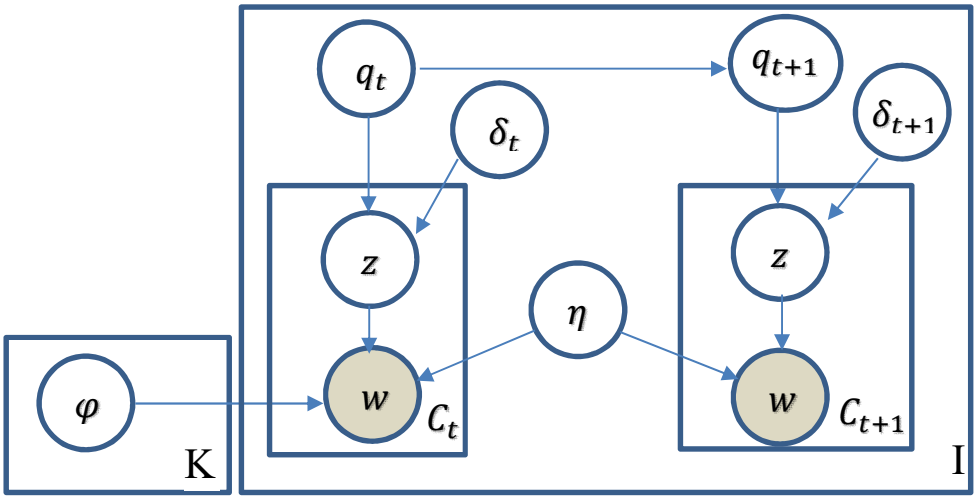

Figure 13 Graphical Model of THMM3

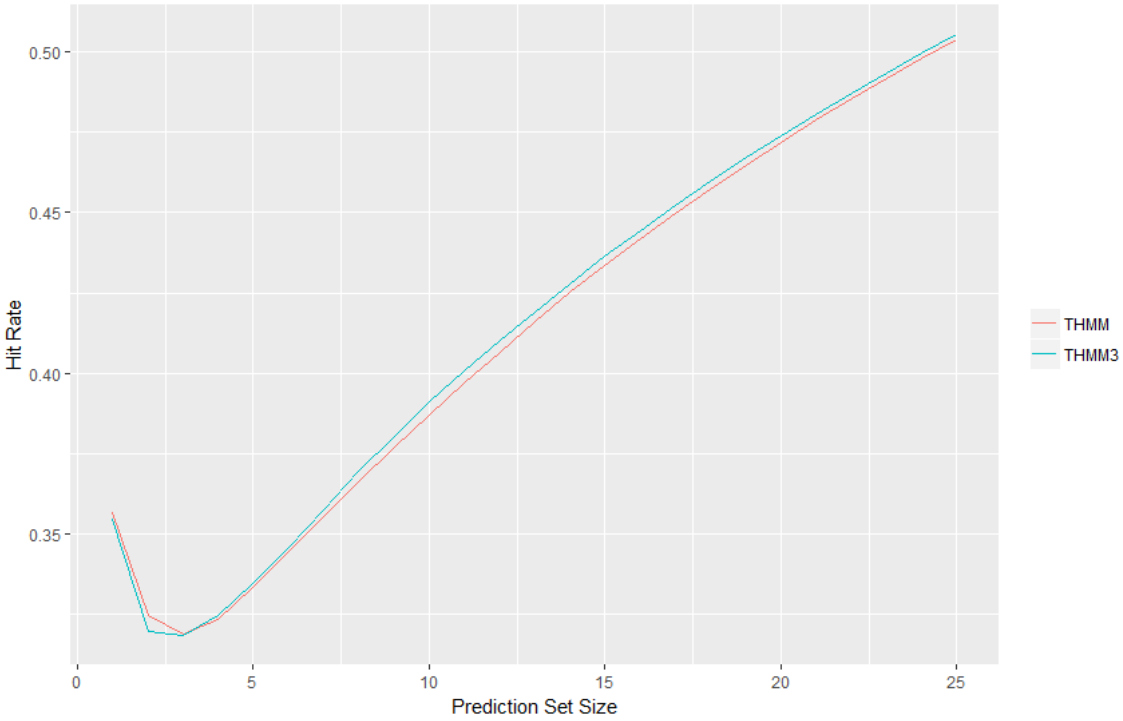

Figure 14 Hit Rate Comparison THMM and THMM3

*THMM4 (Time-varying **δ** and One Static Motivation)*

THMM4 is a combination of THMM2 and THMM3. In other words, I allow for time-varying δ parameters and restrict the model to only one static motivation. This is a

test for improving efficiency and perhaps the accuracy of THMM3. One problem with large number of parameters in THMM3 is that it might lead to overfitting. Therefore, in THMM2, I replace individual-specific parameters with one common static motivation to alleviate this problem.

Figure 15 shows the graphical model of THMM4, and Figure 16 compares its prediction accuracy with that of THMM. These results suggest that dropping the individual-specific motivations is costly for the model.

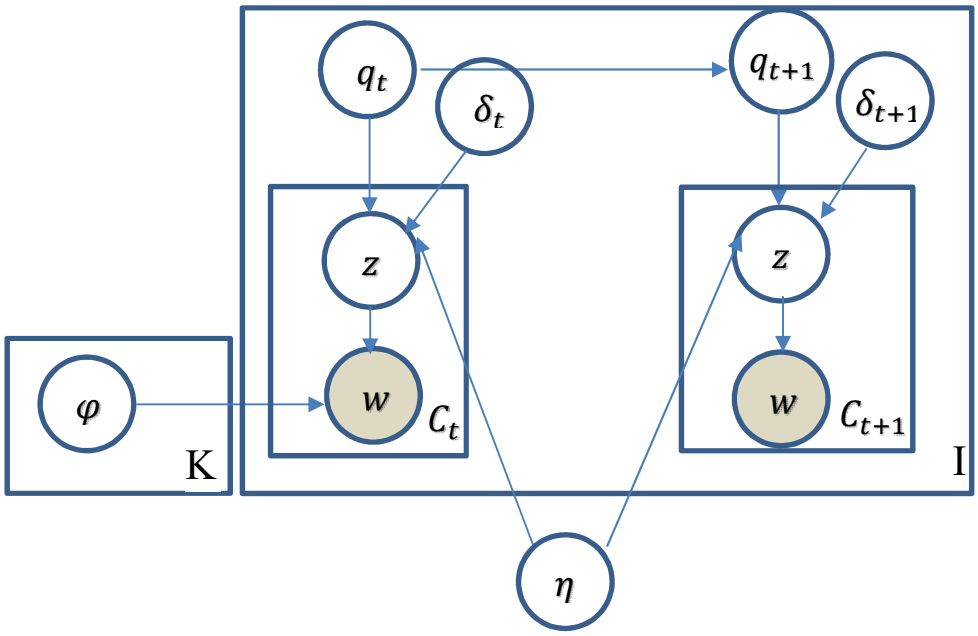

Figure 15 Graphical Model of THMM4

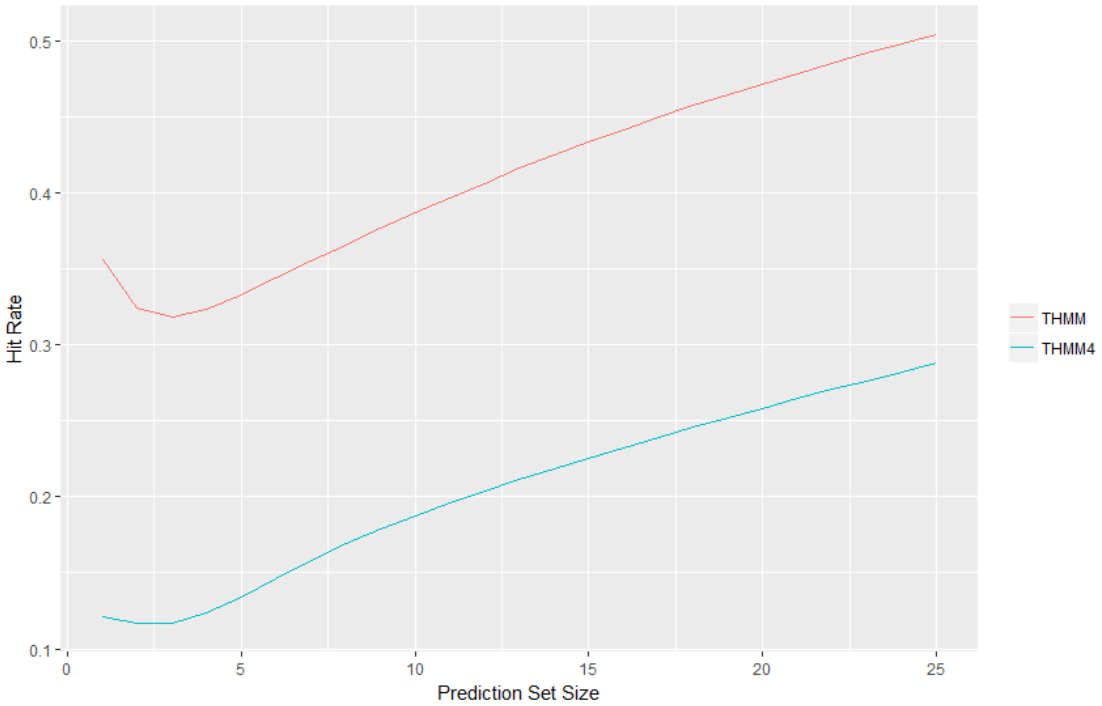

Figure 16 Hit Rate Comparison for THMM and THMM4

In the final analysis, I compare the hit rates of THMM2 and THMM4. Based on my results thus far, I did not expect to see a significant difference. Figure 17 shows the hit rate comparison; for smaller prediction set sizes, THMM4 performs better, but the curves converge as the prediction set size increases. Therefore, I can conclude that the time-varying δ parameter cannot increase the accuracy of the model and will only cause overfitting and slow convergence.

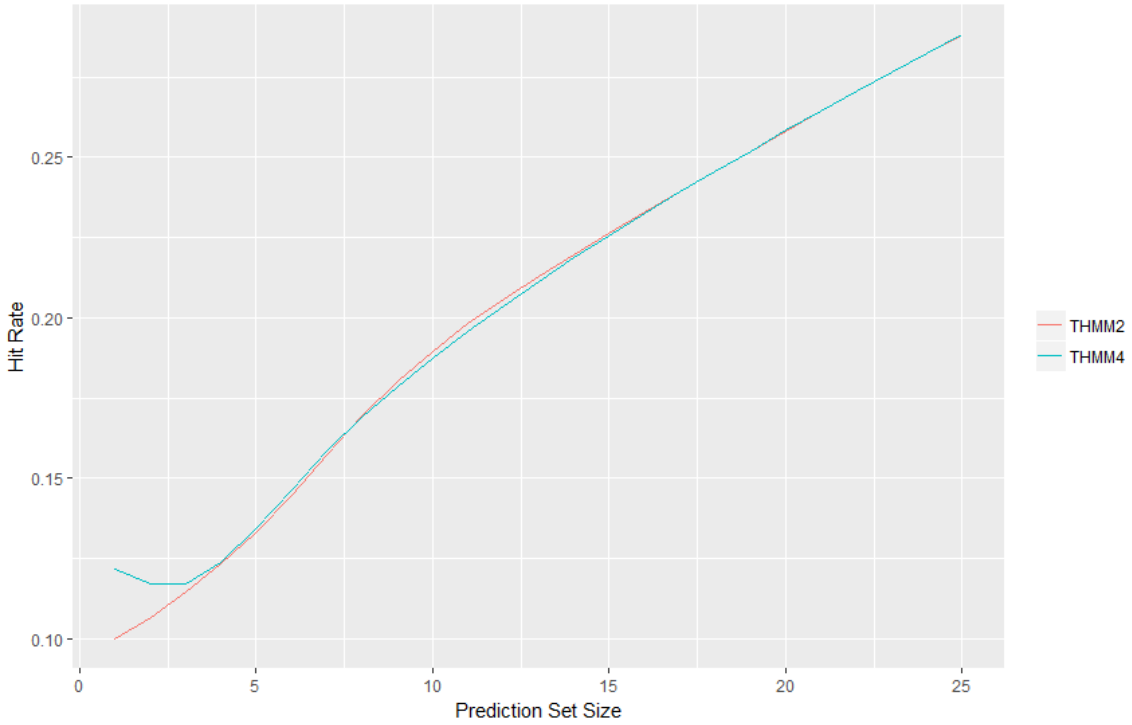

Figure 17 Hit Rate Comparison for THMM2 and THMM4