Segmentation of Financial and Marketing Data:

Mixture Logit Model and Hidden Markov Model

BY

ZIQIAN HUANG
B.A., South China University of Technology, China, 2000
M.S., South China University of Technology, China, 2003
M.S., University of Central Florida, U.S.A., 2005

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Business Administration
in the Graduate College of the
University of Illinois at Chicago, 2011

Chicago, Illinois

Defense Committee:

Stanley L. Sclove, Chair and Advisor
Alan J. Malter
Fangfang Wang
Houston H. Stokes
George Karabatsos, College of Education

To my parents,

my wife: Lei Shu

my daughter: Elizabeth Yile Huang

and my son: Edward Yiqiang Huang

ACKNOWLEDGMENTS

First and foremost I offer my sincerest gratitude to my advisor, Professor Stanley Sclove, who has supported me throughout my dissertation with his patience, suggestions and encouragements. It would have been next to impossible to write this dissertation without Professor Sclove's help and guidance, for example all the datasets in this research are purchased or obtained by Professor Sclove. I always feel very lucky to work with him for six years; from him not only I found my research topic, but also learned the characteristics of a good researcher who treats the research as a joy instead of work.

I am heartily thankful to my dissertation committee: Professor Alan Malter, Professor Fangfang Wang, Professor George Karabatsos and Professor Houston Stokes. They provided many good suggestions to improve the dissertation, and had open minds to my research. They reminded me that this dissertation is just a start point; many opportunities can be explored in my future research.

I would like to thank Dr. Hsing-Chien (Cindy) Kao, Dr. Yongfang Zhu, and Dr. Cuilan Zhang, for helping me to write the dissertation in Latex.

Last I owe my deepest gratitude to my family, especially my wife, Lei. Without their love, understand and supports, no single word can be written in this dissertation.

TABLE OF CONTENTS

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DMEF	Direct Marketing Education Foundation
EM	Expectation-Maximization
FMM	Finite Mixture Model
GLM	Generalized Linear Model
GPR	Gross Product Revenue
HMM	Hidden Markov Model
LRT	Logistic Regression Tree
MLE	Maximum Likelihood Estimation
MLM	Mixture Logit Model
MLR	Mixture Logistic Regression
ROR	Rate of Return
S&P500	Standard and Poor's 500 Composite Stock Index

SUMMARY

Segmentation refers to the assignment of each consumer to a set of similar consumers. The formation of the sets and the assignments are done simultaneously in an algorithm. We focus on utilizing the Mixture Logit Model (MLM) and the Hidden Markov Model (HMM) to segment financial and marketing data. Both the MLM and the HMM originate from the Finite Mixture Model (FMM). The MLM is also called the Mixture Logistic Regression (MLR). Traditionally, cluster analysis, including the K means algorithm and hierarchical methods, have been used as segmentation methods. Cluster analysis is unsupervised learning, in that there is no target variable. In the marketing and finance areas, segmentation results are often considered as one of the important inputs for modeling a target variable, because of the existence of different underlying market segments, both in theory and in reality. Our proposed segmentation methods begin with modeling a target variable, instead of unsupervised learning, and then the existence of segments is evaluated through certain model selection criteria. The characteristics of each segment are shown from their parameter estimates, and further the segments can be profiled by other variables which are not used in the model. This research makes a contribution by illustrating how to segment financial and marketing data objectively and systematically, with regard to incorporating the segmentation into the supervised modeling. We apply the MLM on one marketing solicitation responder dataset, the HMM on the S&P 500 monthly return data, and on one charity donation dataset. All of the results demonstrate that the MLM and the HMM perform better than the benchmark models or methods.

CHAPTER 1

INTRODUCTION

1.1 Research Background

This dissertation deals with applications in marketing research and in financial time series analysis. *Segmentation* will be considered in terms of research on customers and on the stock market.

A market *segment* is a subset of people or organizations sharing one or more characteristics that cause them to have similar product needs. *Market segmentation* is the name given to the set of techniques for forming consumer segments.

In statistics, a *cluster* in a set is a subset of similar objects or persons. *Cluster analysis* is the name given to the set of statistical techniques for forming clusters.

So, cluster analysis is a basis for market segmentation. See, e.g., Churchill and Iacobucci (2004) (12), Kotler and Lane (2009) (27), and Kotler and Armstrong (2010) (26) for general discussions of market segmentation. Segmentation assigns the set of customers into different subsets, and identifies those particular customers in a market on the basis of external variables. Customers in the same segment may respond similarly to a market stimulus, and can likely be reached by a market intervention aimed at their group. The objective of marketing segmentation is to identify groups of similar customers and potential customers, to prioritize the groups to address, to understand their behavior, and to respond with appropriate marketing

strategies that satisfy the different preferences of each chosen segment. Good segments must be measurable, sizable, accessible, actionable, and have competitive intensity and growth potential (30).

A segment of a time series is a sequence of time points in which the observed variables have similar behavior. Segments of time series are also called *regimes*, *epochs* or *phases*. Customers may go through phases. A credit-card customer may be a *transactor* for a period of time, meaning that he or she pays the total balance at the end of the month, or the customer may be a *revolver* for a period of time, meaning that he or she carries over a portion of the balance to the next month. Similarly, donors to a charity or institution may go through donor periods and non-donor periods. An example of time-series segmentation, to be considered here, is the segmentation of stock market prices into Bull and Bear periods.

1.1.1 Labeling and Segmentation Models

For such problems, a class of models which we call *labeling models* is discussed and applied. Each observation Y_i comes with an unobserved *label* S_i which names the state or cluster of case i. In the context of such a model, clustering is estimation of the labels.

A good clustering is homogeneous within segments, but heterogeneous across segments. There should be relatively large between-cluster variation, and relatively small within-cluster variation.

There are *hierarchical* and *non-hierarchical* clustering methods. Hierarchical clustering is either divisive or agglomerative. Agglomerative hierarchical clustering starts with each case in its own cluster, and then puts the two closet individuals together, then the next closest, etc.

Divisive clustering starts with all cases in a single cluster, and then breaks the weakest link, then the next weakest, etc. K-means partitioning separates the dataset into a prespecified number of segments, and then reallocates or swaps customers to improve some measure of effectiveness. In the name K-means, the symbol K denotes the number of segments, and "means" refers to the fact that the algorithm proceeds by assigning each case to the nearest of a pre-specified set of seed points ("means"), then updates the seeds, then re-assigns the cases, etc. For a fixed number K of segments, usually two methods will come up with similar results for the same data. Different values of K can be tried and the results compared with a model selection criterion.

Clusterwise regression can be used to explore the relationship between response variables and explanatory variables, within segments. A given explanatory variable can have a different relationship to the response variable in one segment than it does in another. The sign of its coefficient might even be positive in one segment and negative in another. The Logit model can be used when the response is binary or categorical. Of course, the Logit model is widely used for the classification problem, in all kinds of research and applications. In marketing research, the Logit model for customer choice helps analyze and explain the choices individual customers made in a market. It enables firms to understand the extent to which factors influence a customer's choice, such as buying or not buying a product, responding or not responding to a direct mail solicitation, etc. Firms also can use customer choice analysis to develop marketing programs tailored to specific market segments, or even to individual customers. When the Logit model is used clusterwise, we call the model the Mixture Logit Model (MLM).

1.1.2 Research Motivation

The relationships between marketing segmentation and classification depend on the marketing research objective. Given that segmentation is the main purpose, a marketer can use the Logit model to obtain each customer's probability of a particular decision (for example, the probability that the customer will buy the product), and segment the customers based on those probabilities. If understanding a customer's choice is the goal, sometimes trying to segment the set of customers first is a necessary step, because in different segments there might be different reasons for that choice. In our research, we still consider understanding a customer's behavior as the goal, but how to incorporate the segmentation into this process is an interesting question. From the modeling perspective, usually segmentation is unsupervised learning in that there is no specific target variable; the data are naturally grouped according to some criterion, such as within cluster sum of squares for K-means clustering. The existence of a segment in a classification model is proven either by different variables being important in different segments or by parameter estimates being different across segments. The above sequential procedure is feasible in practice in that it is easy to perform and explain. The nature of unsupervised learning raises our concerns about such a procedure. Since at first the segmentation process is not specific to a target variable, it is questionable whether the segmentation results benefit the modeling of the target variable, which randomly depends on a relationship to the segmentation results. Logically speaking, even if the segmentation variable is not significant in the model, or parameter estimates are similar in different segments, it is futile to deny the existence of underlying segments. Maybe segmenting data in another way will show the existence of under-

lying segments in the model, such as by changing the predefined number of segments or the segmentation algorithm. Unfortunately there is no systematic way to decide when to start and stop such a procedure, and it is done in ad-hoc style. An improved solution is to do the two tasks of modeling the target variable and performing the segmentation simultaneously. In that way, the segmentation is switched from unsupervised to supervised learning. In unsupervised learning, it is as if the segmentation serves to define a target variable whose values correspond to the labeling.

We use two extensions of the FMM. One is the MLM, which is clusterwise regression, in particular, clusterwise logistic regression. The second extension is an extension to time series, the HMM. Both the MLM and the HMM (Hidden Markov Model) can be used to segment and classify the data simultaneously. The MLM is the combination of the Logit model and the FMM (Finite Mixture Model). The MLM is same as the Logit model in that it models the success probability p through the Logit link function, while allows different Binomial distributions with different values of p. That is, the overall distribution can be the result of the mixture of several Binomial distributions. To put it another way, the MLM relaxes a major assumption of the Logit model, that there is only one component, with its value of p. The population of origin of each observation can be viewed as its segment. The HMM is another extension of the FMM, since in addition to treating the data as coming from a mixture distribution, it allows data component membership to change from time to time, following a Markov process. Especially when the data are longitudinal, the HMM might be a good choice given that there are some underlying hidden states, and that the component membership jumps among the different states.

The HMM is a dynamic FMM. A time series is conceived as moving back and forth through K states. The process can be represented as (Y_t, S_t), $t = 1, 2, \ldots, n$, where t represents time, Y_t is observed, and S_t is the unobserved state at time t, $S_t = 1, 2, \ldots, K$ states. Different number of states, that is, different values of K, can be tried and the results can be compared with model selection criteria. The elements of the HMM are the class-conditional probability density functions (p.d.f.) $f_k(y)$, $k = 1, 2, \ldots, K$, and transition probabilities $p_{jk} = \Pr\{S_t = k \,|\, S_{t-1} = j\}$, $j, k = 1, 2, \ldots, K$. The FMM has a p.d.f. in terms of mixing probabilities λ_k, $k = 1, 2, ..., K$,

$$f(y_i) = \lambda_1 f_1(y_i) + \lambda_2 f_2(y_i) + \cdots + \lambda_K f_K(y_i),\ i = 1, 2, \ldots, n.$$

The HMM can be described in a parallel manner, with transition probabilities replacing mixing probabilities in the state-conditional p.d.f.s, as

$$f(y_t \,|\, S_{t-1} = j) = p_{j1} f_1(y_t) + p_{j2} f_2(y_t) + \cdots + p_{jK} f_K(y_t),\ t = 1, 2, \ldots, n.$$

1.2 Literature Review

This following review is not intended to be comprehensive or bibliographic but rather mentions some particular works in the development of the models to be considered here.

1.2.1 Clustering Algorithms

A *cluster* may be defined as a set of similar cases. A *clustering algorithm* is a procedure for separating a sample into clusters. Among early clustering algorithms are ISODATA (Ball

and Hall 1965) and K-means (MacQueen 1967). Here K denotes the number of clusters to be fit. One can of course try different values of K. Both algorithms ISODATA and K-means start with K initial seeds, working through the sample case by case, assigning each case to the seed point to which it is closest. The choice of initial seeds can make a difference, and usually one tries clustering the data several times with different sets of initial seeds. The initial seeds may be chosen at random, or by design. ISODATA makes a pass through all n observations, while K-means updates the seeds after each case is assigned. SAS procedure FASTCLUS is an implementation of K-means; it has an OPTION called DRIFT, which implements the updating of the K-means after each case is assigned, rather than waiting until all n have been assigned.

There were many papers on clustering algorithms scattered throughout the literature of various fields, and it was a help when books appeared. An early major book on clustering is Hartigan's (1975) book (21), which focuses on the description of algorithms. An even earlier book is Anderberg's (1974)(1).

1.2.2 Model-based Clustering

As soon as one invokes a concept of "closeness", "similarity", "distance" or "dissimilarity", the question of what may be an appropriate metric arises. This points to the need for a model.

John (1969) considered the problem of identifying the population of origin of each observation in a sample that is from a mixture of two Normal distributions (23). He also (1970) considered a similar problem for a mixture of two Gamma distributions (24). Wolfe (1970) considered pattern clustering by multivariate mixture analysis (50). Sclove (1977) considered the problem of clustering individuals in the context of a mixture of $K \geq 2$ distributions, and

showed there are relationships between Ball and Hall's "isodata" (1965) procedure and MacQueens (1967) K-means procedure, in that both work from seed points, ISODATA updating the seeds after each case is assigned, whereas K-means loops through the whole dataset before updating the seeds. Sclove showed that these algorithms can be viewed as iterative methods of maximum-likelihood estimation in a mixture model. He focused on the classification likelihood, in which the labels are considered as already fixed. In the mixture likelihood the labels are random, with a multinomial distribution with K categories with probabilities λ_k, $k = 1, 2, \ldots, K$. Once the sampling is done, the labels are a realized sample from this multinomial distribution (43). Another modification puts a prior on K.

Books on model-based clustering clustering via the Finite Mixture Model (FMM) began to appear in the 1980s; these include Everitt and Hand (1981) (16) and Titterington, Markov and Smith (1985) (46). The book by McLachlan and Basford (1988) highlights the importance of finite mixture distributions in modeling heterogeneous data, with a focus on applications in cluster analysis. Their emphasis is on maximum likelihood estimation, via the EM algorithm (35). An updated edition is McLachlan and Peel (2000) (36).

An important aspect of clustering is the *profiling* of the samples within the clusters, particularly with respect to the cluster-specific means of the variables. The set of profiles can be exhibited in a two-way table of means, variables by clusters. "Segmentation" really has come to refer not just to the clusters but to the process of profiling and studying the clusters obtained. The FMM is a basis for incorporating such segmentation into classification.

1.2.3 Clusterwise Regression

Often the variables are divided into two subsets, variables to be explained or predicted (response, or dependent variables), and variables to use for this (explanatory, or independent variables). Hence, *clusterwise regression,* regression within clusters, becomes important.

DeSarbo and Cron (1988) presented a conditional mixture, maximum likelihood methodology for performing clusterwise linear regression. Their new methodology simultaneously estimates separate regression functions and membership in K clusters or groups (14).

Leung (2001) proposed a regression-class mixture decomposition method for the mining of regression classes in large datasets. A new concept, called "regression class" which is defined as a subset of the dataset that is subject to a regression model, is proposed as a basic building block on which the mining process is based on regression (28). Their work is for linear regression.

Grün and Leisch (2006) developed a package flexmix in R, which provides flexible modeling of finite mixtures of regression models using the EM algorithm. The mixture can be a mixture of logistic regressions, the corresponding model being called Mixture Logistic Regression (MLR) or the Mixture Logistic Model (MLM). The use of the software in addition to model selection is demonstrated on a Logistic regression example (19).

Deodhar (2007) also mentioned that for difficult classification or regression problems, practitioners often segment the data into relatively homogenous groups and then build a model separately for each group. They proposed a framework generalized co-clustering and collaborative filtering to model-based co-clustering, which can be viewed as simultaneous co-segmentation and classification or regression (13). Their work is not based on the FMM.

1.2.4 Hidden Markov Model

The concept of a labeling model can be carried over into the domain of time series analysis. The *hidden Markov model* is a labeling model that is a dynamic extension of the FMM. In a dynamic labeling model, the vector (Y, S) is an element of a time series $\{(Y_t, S_t), t = 1, 2, \ldots, n\}$. In the HMM, the labels constitute a Markov chain. Thus, there are probabilities on the transitions between states. What state is likely at time t depends upon the state at time $t - 1$.

First used for speech recognition (Rabiner 1989) (41), HMMs are now applied in a variety of disciplines, including such different areas as marketing, medicine, and sports.

Medicine. Scott, James and Sugar (2005) used an HMM to address the combination of clustering and longitudinal analysis, and compared the effectiveness of clozapine and haloperidol, two antipsychotic medications for schizophrenia.They compared a two-stage procedure of clustering followed by estimation of transition probabilities with a single-stage procedure of HMM estimation which more properly iterates back and forth between the two. They found better results with the single-stage rather than the two-stage procedure (45).

Marketing. When the transition probabilities of a Markov chain are time-stationary, the process is said to be *homogeneous.* Otherwise, as when the transition probabilities are not stationary but rather functions of time-varying covariates, the process is said to be *nonhomogeneous.* Netzer, Lattin and Srinivasan (2008) constructed and estimated a nonhomogeneous HMM to model the transitions among latent relationship states and effects on buying behavior (40).

Sports. Jensen, McShane and Wyner (2009) used HMM to predict the hitting performance of major league baseball players, and compared the model to current sabermetric methods on a hold-out season (22). (*Sabermetrics* is the analysis of baseball through objective, empirical evidence, especially baseball statistics that measure in-game activity rather than industry activity such as attendance. The term is derived from the acronym SABR, which stands for the Society for American Baseball Research. It was coined by Bill James, one of the founders of SABR, often considered its most prominent advocate.)

1.3 Research Agenda

Because they accomplish classification and facilitate profiling in the content of a mixture and a dynamic extension of the mixture models, the MLM and HMM are our models of interest. They are used to analyze one or another of three datasets. In Chapter Two, we introduce the MLM, and apply it on one marketing solicitation responder dataset from the Direct Marketing Education Foundation (DMEF). In Chapter Three, we further discuss the HMM, and in a financial application we segment S&P500 monthly Rate of Return (ROR) into different states. In Chapter Four, we return to the marketing setting and further study a charity donation dataset from the DMEF, based on the HMM.

These three chapters can be read independently, but the introductions of the MLM and the HMM are in Chapter Two and Chapter Three respectively. While the segmentation concept and the model complexity are increasing from Chapter Two to Chapter Four. Chapter Two is a starting point for demonstrating how to integrate the segmentation with the classification, by extending the FMM to the MLM. Chapter Three further extends the FMM to the more compli-

cated HMM by allowing the dynamic transition between the segments, while we only consider one HMM sequence without any predictors in the model. Chapter Four involves two thousand independent HMM sequences and one independent variable to model the transition probability matrix, which requires more intensive computation, but is more practical for marketing longitudinal data segmentation.

CHAPTER 2

A SEGMENTATION FOR A SOLICITATION RESPONDER DATASET BASED ON THE MIXTURE LOGIT MODEL

2.1 Introduction

We are interested in the Mixture Logit Model (MLM), which is the combination of the Finite Mixture Model (FMM) and the Logit model. The Logit model belongs to the Generalized Linear Model (GLM) family. The GLM family is characterized by link function $g(.)$ which link the regression function $\mathcal{E}[Y|\boldsymbol{x}]$ to a linear model $g(\mathcal{E}[Y|\boldsymbol{x}]) = \boldsymbol{\beta}'\boldsymbol{x}$. The GLM is widely used in current research and industrial practice for the classification problem. Marketing segmentation is a common practice in the marketing research, based on the fact that different customers have different needs and wants because of different attitudes, life stage, etc. We want to use the MLM to fit data usually modeled by the Logit model, to assess the value of the MLM, more importantly we try to segment a marketing dataset following the framework of the MLM, to provide a new method for marketing segmentation.

In this chapter, we first review the Finite Mixture Model, the GLM, the Logit model and the MLM. Then we perform simulation studies on two datasets to illustrate the necessity of the MLM. Finally, we analyze a marketing dataset from the Direct Marketing Education Foundation (DMEF), using the MLM to model it, and propose a marketing segmentation.

2.2 The Finite Mixture Model (FMM)

2.2.1 Notation for FMM

The following is notation for the FMM.

$\boldsymbol{y}$, vector of p variables

$f(\boldsymbol{y})$, probability density function (p.d.f.)

(If $\boldsymbol{y}$ is discrete, we call this a probability mass function (p.m.f.).)

n number of cases (sample size)

$i = 1, 2 \ldots, n$ cases

$\boldsymbol{y}_i$, observation vector for the i-th case

K number of states

$k = 1, 2, \ldots, K$ states

Π_k, the k-th population (segment or component)

S_i, state of case i; equal to 1, 2, $\ldots, k, \ldots,$ or K. The event $S_i = k$ is equivalent to the event that case i originated from Π_k, that is, the population of origin of case i is Π_k.

$f_k(\boldsymbol{y})$, the state-conditional p.d.f., that is, conditional p.d.f., given state k: $f_k(y) = f(y_i|S_i = k)$

2.2.2 Model for FMM

The p.d.f. is of the form

$$f(\boldsymbol{y}_i) = \lambda_1 f_1(\boldsymbol{y}_i) + \lambda_2 f_2(\boldsymbol{y}_i) + \cdots + \lambda_k f_k(\boldsymbol{y}_i) + \cdots + \lambda_K f_K(\boldsymbol{y}_i) = \sum_{k=1}^{K} \lambda_k f_{ki},$$

where $f_{ki} = f_k(y_i)$, $\sum_{k=1}^{K} \lambda_k = 1$, and $\lambda_k > 0$. The λ_k are the *mixing probabilities* and the $f_k(\cdot)$ are the *component*, or *class-conditional* or *state-specific* densities.

For a mixture of regressions, the component densities are to be conditional densities for Y given $\boldsymbol{x}$.

2.2.3 Profiling for FMM

After segmentation, mean profiling is accomplished by displaying a table of estimates of the segment means. Note that $\boldsymbol{Y}$ is in general a vector of p variables; then the profile of segment k is the vector of p estimated means. The set of profiles across the K segments is then a variable-by-segment $(p \times K)$ table. For $k = 1, 2, \ldots, K$, the estimate of the mean (or mean vector) of segment k is

$$\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n} w_{ki} \boldsymbol{y}_i / \sum_{i=1}^{n} w_{ki}, \tag{2.1}$$

where for soft classification

$$w_{ki} = \Pr(\, \Pi_k \,|\, \boldsymbol{y}_i \,)$$

and for hard classification

$$w_{ki} = 1 \text{ if } k = \arg \max_k \{ \Pr(\, \Pi_k \,|\, \boldsymbol{y}_i \,) \} \text{ and } w_{ki} = 0 \text{ otherwise.}$$

Then $\sum_{i=1}^{n} w_{ki} = n_k$, the number of cases classified as segment k, and the statistic in the profile is the ordinary sample mean of those cases.

Note that the posterior probability of membership of Case i in segment k is

$$\Pr(\,\Pi_k \,|\, y_i\,) = \lambda_k f_k(y_i)/f(y_i) = \lambda_k f_k(y_i)/\sum_{j=1}^{K} \lambda_j f_j(y_i). \tag{2.2}$$

2.3 Logit Model

The *generalized linear model* for a scalar response variable Y is

$$g(\mathcal{E}[Y|\boldsymbol{x}]) = \boldsymbol{\beta}'\boldsymbol{x},$$

where $g(\cdot)$ is the *link function*, i.e., function that links $\mathcal{E}[Y|\boldsymbol{x}]$ to a linear model in $\boldsymbol{x}$ and

$$\boldsymbol{x}' = (1, x_1, x_2, \ldots, x_p), \boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p).$$

When Y is binary $(0, 1)$, its expectation is $\Pr(Y = 1) = \pi$, where $0 < \pi < 1$. The *odds* are $\text{Odds}(\pi) = \pi/(1-\pi), 0 < \text{Odds}(\pi) < \infty$. The conditional expectation $\mathcal{E}[Y|\boldsymbol{x}]$ is $Pr[Y = 1|\boldsymbol{x}]$, often denoted simply by a symbol such as P_x or π_x.

The *logit* is the link function

$$g(\pi) = \text{logit}(\pi) \;=\; \log\,\text{Odds}(\pi) \;=\; \ln[\pi/(1-\pi)] = \ln\pi - \ln(1-\pi),$$

where $0 < \pi < 1$, $0 < odds < \infty$, and $-\infty < \text{logit} < \infty$. Let $z = \text{logit}(\pi)$. Then the inverse of the logit function is the *logistic function,*

$$h(z) = g^{-1}(z) = e^z/(1 + e^z) = 1/(1 + e^{-z}),\ 0 < h(z) < 1.$$

It is an example of a *squashing function*, mapping the real line into (0, 1).

In logistic regression, the logit is regressed on the explanatory variables. The inverse $h(\boldsymbol{\beta}'\boldsymbol{x}) = g^{-1}(\boldsymbol{\beta}'\boldsymbol{x})$ of the logit link function is, letting $z = \boldsymbol{\beta}'\boldsymbol{x}$,

$$h(\boldsymbol{\beta}'\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x})} = \frac{1}{1 + \exp(-\boldsymbol{\beta}'\boldsymbol{x})}.$$

The data are $\boldsymbol{x}_i, y_i,\ i = 1, 2, \ldots, n$. The likelihood is

$$L = \prod_{i=1}^{n} f_i = \prod_{i=1}^{n} p_i^{y_i}\ q_i^{1-y_i},$$

where

$$f_i = p_i^{y_i}\ q_i^{1-y_i},\ p_i = h(\boldsymbol{\beta}'\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)} = \frac{1}{1 + \exp(-\boldsymbol{\beta}'\boldsymbol{x}_i)}.$$

and

$$q_i = 1 - p_i = 1 - h(\boldsymbol{\beta}'\boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)} = \frac{\exp(-\boldsymbol{\beta}'\boldsymbol{x}_i)}{1 + \exp(-\boldsymbol{\beta}'\boldsymbol{x}_i)}.$$

2.4 Mixture Logit Model

The regression coefficients are in general different in different segments. Therefore, it is of interest to profile the segments in terms of the coefficients of the explanatory variables. The model is of the form

$$g(\mathcal{E}[Y|\boldsymbol{x}, \Pi_k]) = \boldsymbol{\beta}_k'\boldsymbol{x}.$$

where $g(\cdot)$ is the link function.

For logistic regression, this is

$$\text{logit}(\pi_k) = \boldsymbol{\beta}_k'\boldsymbol{x},$$

where

$$\pi_k = Pr(Y = 1|\boldsymbol{x}, \Pi_k).$$

The probability mass function (p.m.f.) of the k-th component is

$$f_k(y) = f(y, \beta_k), y = 0, 1.$$

Index the components by $k = 1, 2, \ldots, K$ and the observations by $i = 1, 2, \ldots, n$. Let

$$p_{ki} = \frac{\exp(\boldsymbol{\beta}_k'\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}_k'\boldsymbol{x}_i)} = \frac{1}{1 + \exp(-\boldsymbol{\beta}_k'\boldsymbol{x}_i)}.$$

and

$$q_{ki} = 1 - p_{ki} = \frac{1}{1 + \exp(\boldsymbol{\beta}_k'\boldsymbol{x}_i)} = \frac{\exp(-\boldsymbol{\beta}_k'\boldsymbol{x}_i)}{1 + \exp(-\boldsymbol{\beta}_k'\boldsymbol{x}_i)}.$$

Let

$$f_{ki} \;=\; p_{ki}^{y_i}\, q_{ki}^{1-y_i}.$$

The p.m.f. of Y_i is

$$f_i = \sum_{k=1}^{K} \lambda_k f_{ki}.$$

The likelihood is

$$L = \prod_{i=1}^{n} f_i.$$

It may seem to be a bit of a puzzle how components can be needed, when there are only two values for Y. But, the differences among the components k are in the relationships of the logit to the predictors, via the coefficient vectors $\boldsymbol{\beta}_k, k = 1, 2, \ldots, K$. It can still be interesting to segment the customers, and to examine the profiles of the segments.

2.4.1 Method to Estimate the MLM

A method to estimate the MLM is maximum likelihood estimation, namely we try to find a set of parameter values which maximize the likelihood function of the MLM. A Newton-Raphson numerical method can be utilized, and an Expectation-Maximization (EM) algorithm can also be used. The following EM algorithm might be employed to estimate the MLM. We put the details of the derivation in the Appendix.

- Initialize the probabilities $p(k|y_i, \boldsymbol{x}_i)$ at $p^{(0)}(k|y_i, \boldsymbol{x}_i)$. For example, these could all be chosen as $1/K$.

- Numerically solve

$$\sum_{i=1}^{n} x_{vi}\, p^{(0)}(k|y_i, \boldsymbol{x}_i)\, y_i \;=\; \sum_{i=1}^{n} x_{vi}\, p^{(0)}(k|y_i, \boldsymbol{x}_i)\, \{1/[1+\exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)]\}$$

  for $\boldsymbol{\beta}_k$.

- Update f_i.

- Update p_{ki}.

- Update $p(k|y_i, \boldsymbol{x}_i)$ and λ_k, $k = 1, 2, \ldots, K$.

- Update β_k.

- Etc., until satisfactory convergence.

Leisch and Friedrich (2006) developed a package called 'FlexMix' in R to estimate the MLM, where they used an EM algorithm. After reviewing their recent publications, and testing the software performance, we decide to use their package to estimate the MLM in this research (19).

2.4.2 Reasons for Using the Mixture Logit Model

2.4.2.1 Simulation Objective and Parameter Settings

The MLM does not require that the given dataset comes from one Binomial distribution as the Logit model assumes. It is necessary to investigate this one-distribution assumption impact on those data which are generated from different Binomial distributions. For this purpose, we adopt a data simulation method here.

We simulate two datasets here, dataset A and B, and for simplicity, both datasets only contain the samples from two different Binomial distributions, with one intercept and two predictors for the logit function of p. The difference between datasets A and B is that two predictors should be significant in the model estimate in A, while in B only one of the two predictors is actually used to simulate sample data.

The parameters of dataset A and B are listed below:

Dataset A:

Sample size $n = 10,000$,

The number of predictors $p = 2$,

Predictor 1: $X_1 \sim Uniform(0, 1)$,

Predictor 2: $X_2 \sim Uniform(0, 1)$,

Component 1 Model: $Logit(p_1) = 1 - 4x_1 + 2x_2$,

Component 2 Model: $Logit(p_2) = -20 + 80x_1 - 40x_2$,

Mixing probability: $\pi_1 = 0.4$ and $\pi_2 = 0.6$.

Dataset B:

Sample size $n = 10,000$,

The number of predictors $p = 2$,

Predictor 1: $X_1 \sim Uniform(0, 1)$,

Predictor 2: $X_2 \sim Uniform(0, 1)$,

Component 1 Model: $Logit(p_1) = 1 - 2x_1$,

Component 2 Model: $Logit(p_2) = -20 + 40x_2$,

Mixing probability: $\pi_1 = 0.5$ and $\pi_2 = 0.5$.

When we generate those datasets, we also create a field called *cluster*(*label*) to indicate the population origin of each case. The cross-tabulation of the cluster and the simulated Y of dataset A is Table I, and the counterpart of dataset B is Table II. There are balanced portions of 0's and 1's in data set A and B.

TABLE I

CROSS-TABULATION OF THE CLUSTER AND Y FROM DATASET A

Y	Cluster	
	1	2
0	2,023	2,976
1	1,979	3,022

2.4.2.2 Models for the Simulated Datasets

2.4.2.2.1 Logit Model with Given Cluster Labels

We first try to fit the datasets A and B with the Logit model by cluster, assuming that we know each data point belongs to which population, or in other words, the *cluster* variable is known in the datasets. Using the 'glm' package in R, we obtain the Logit model estimate for each dataset. The software output for the component 1 Logit model of the dataset A is

TABLE II

CROSS-TABULATION OF THE CLUSTER AND Y FROM DATASET B

Y	Cluster	
	1	2
0	2,410	2,549
1	2,591	2,450

Figure 1, and the output for the component 2 Logit model of the dataset A is Figure 2. The software output for the component 1 Logit model of the dataset B is Figure 3, and the output for the component 2 Logit model of the dataset B is Figure 4.

The results obviously show that when we know the data from two different distributions, and the data cluster membership is known, fitting separate Logit models for each cluster can obtain good estimations of true parameters.

2.4.2.2.2 Logit Model with Unknown Clusters

In reality, sometimes or most of time the *cluster* variable is unknown, or there is no such perfect *cluster* variable to indicate data point origin. The common practice is either to assume the whole data comes from one distribution, or to segment the whole data based on some variables first, and then fit a Logit model within each segment. Here we do not have other variables to segment data, therefore we just work on what happens if we fit traditional one-component Logit models on datasets A and B.

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3215  -0.8704  -0.3481   0.8709   2.4226

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.11574    0.09613   11.61   <2e-16 ***
x1          -4.14435    0.14813  -27.98   <2e-16 ***
x2           1.88664    0.13444   14.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5547.5  on 4001  degrees of freedom
Residual deviance: 4360.2  on 3999  degrees of freedom
AIC: 4366.2

Number of Fisher Scoring iterations: 4
```

Figure 1. The output for dataset A component 1 Logit model

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data_2)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.393   0.000   0.000   0.000   2.801

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -22.717      1.532  -14.83   <2e-16 ***
x1            91.331      6.066   15.05   <2e-16 ***
x2           -45.951      3.083  -14.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8314.64  on 5997  degrees of freedom
Residual deviance:  441.82  on 5995  degrees of freedom
AIC: 447.82

Number of Fisher Scoring iterations: 11
```

Figure 2. The output for dataset A component 2 Logit model

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data_1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.667  -1.109   0.795   1.064   1.602

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.11513    0.07843  14.217   <2e-16 ***
x1          -1.91486    0.10361 -18.481   <2e-16 ***
x2          -0.16800    0.10083  -1.666   0.0957 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6926.3  on 5000  degrees of freedom
Residual deviance: 6557.7  on 4998  degrees of freedom
AIC: 6563.7

Number of Fisher Scoring iterations: 4
```

Figure 3. The output for dataset B component 1 Logit model

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data_2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.05252  -0.01241  -0.00010   0.00904   2.77163

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.4950     0.9724 -20.049   <2e-16 ***
x1           -0.0165     0.2981  -0.055    0.956
x2           39.1661     1.9247  20.349   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6928.12  on 4998  degrees of freedom
Residual deviance:  849.88  on 4996  degrees of freedom
AIC: 855.88

Number of Fisher Scoring iterations: 9
```

Figure 4. The output for dataset B component 2 Logit model

Using the 'glm' package in R, and fitting the datasets A and B with Logit models, we can get the output for dataset A in Figure 5, and the one for dataset B in Figure 6. From the results, we can find the estimates are far away from the true parameters, in terms of both magnitude and direction. For example, the intercept of the dataset A, for the cluster 1 is 1, and for the cluster 2 is -20, while the estimate of the intercept is -0.67786. The coefficients of $X1$ and $X2$ are all significant in the dataset B Logit model, while from the simulation, we know the coefficient of $X2$ should be insignificant for the data from the cluster 1, and the coefficient of $X1$ should be insignificant for the data from the cluster 2.

In summary, it shows the parameter estimates are biased when fitting the data from the mixture Binomial distribution with the Logit model.

2.4.2.2.3 Mixture Logit Model for the Datasets

In this session, we apply the MLM on the datasets A and B, to assess its performance on parameter estimates, component membership recovery, and prediction capability. The software we used for fitting the MLM is the package 'flexmix' in R.

We choose the number of components to be 2 for both datasets based on the values of AIC, and BIC, which is Figure 7 for the dataset A, and Figure 8 for the dataset B. After choosing the number of components, we get the parameter estimates, the results of the dataset A is in Table III, and the one of the dataset B is in Table IV. From the results, we can see the parameter estimates are much closer to the true ones than the ordinary Logit model ones, in terms of both the sign and the magnitude. One common phenomena, which is called

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7947  -1.0614   0.6697   1.0664   1.7999

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.67786    0.05527  -12.27   <2e-16 ***
x1           2.09428    0.07585   27.61   <2e-16 ***
x2          -0.75420    0.07318  -10.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13863  on 9999  degrees of freedom
Residual deviance: 12939  on 9997  degrees of freedom
AIC: 12945

Number of Fisher Scoring iterations: 4
```

Figure 5. The output for dataset A Logit model

```
Call:
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = sim_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1488  -0.9529   0.4929   0.9416   2.1266

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.05033    0.05725  -18.35   <2e-16 ***
x1          -1.15805    0.07775  -14.89   <2e-16 ***
x2           3.29539    0.08349   39.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13862  on 9999  degrees of freedom
Residual deviance: 11812  on 9997  degrees of freedom
AIC: 11818

Number of Fisher Scoring iterations: 4
```

Figure 6. The output for dataset B Logit model

label switching, appears in the dataset B result. The estimates for the component 2 is for the parameters in the cluster 1.

It is also necessary and interesting to evaluate the MLM capability for identifying the unknown cluster membership. Following the Equation 2.2, we can get each data point's component membership probability. If we assign each data point to the component with the maximum membership probability, it is the "hard" segmentation. For the "soft" segmentation, we can look at the membership probability mean with each cluster. The cross-tabulation between the hard segmentation and the cluster for the dataset A is Table V, and the one for the dataset B is Table VI. The hit rates of the segment are 88.26% and 73.93% for the datasets A and B respectively, which are all larger than the 50% based on random guessing. The average probability in each segment by the cluster for the dataset A is Table VII, and for the dataset B is Table VIII. There are also distinct differences among the mean segment probabilities of different clusters.

From the forecasting aspect, sometimes the model with biased estimates is not a problem, compared to some impact analysis or sensitivity analysis. Therefore, we generate an additional 10,000 data points for the dataset A and B respectively, and those data are not used for model building, but just for model out of sample prediction evaluation. For the dataset A, the hit rate of the Logit model is 68.26%, and the MLM one is 69.53%; For the dataset B, the hit rate of the Logit model is 71.54%, and the MLM one is 73.18%. The result shows the MLM can increase the hit rate to some extent.

```
Call:
stepFlexmix(cbind(y, abs(y - 1)) ~ x1 + x2, data = sim_data,
    control = list(iter.max = 1000, verbose = 100), model = Model1,
    k = 1:3, nrep = 3)

  iter converged k k0    logLik      AIC      BIC      ICL
1    2      TRUE 1  1 -6469.392 12944.78 12966.41 12966.41
2  142      TRUE 2  2 -6066.082 12146.16 12196.64 15326.36
3  359      TRUE 3  3 -6064.383 12150.77 12230.08 17799.16
```

Figure 7. Values of model selection criteria for dataset A MLM

```
Call:
stepFlexmix(cbind(y, abs(y - 1)) ~ x1 + x2, data = sim_data,
    control = list(iter.max = 1000, verbose = 100), model = Model1,
    k = 1:3, nrep = 3)

  iter converged k k0    logLik      AIC      BIC      ICL
1    2      TRUE 1  1 -5905.983 11817.97 11839.60 11839.60
2  108      TRUE 2  2 -5708.580 11431.16 11481.63 17895.36
3  125      TRUE 3  3 -5708.565 11439.13 11518.44 22168.21
```

Figure 8. Values of model selection criteria for dataset B MLM

TABLE III

THE MLM PARAMETER ESTIMATES OF THE DATASET A

	Comp1	Comp2
π	0.424	0.576
Intercept	0.953441	-24.89762
Coef of X_1	-3.306421	98.55608
Coef of X_2	1.419206	-49.40787

TABLE IV

THE MLM PARAMETER ESTIMATES OF THE DATASET B

	Comp1	Comp2
π	0.502	0.498
Intercept	-17.9325058	1.1148037
Coef of X_1	0.2988837	-1.9808339
Coef of X_2	36.0060054	-0.1957188

2.4.2.3 Simulation Results Summary

From two datasets simulation study, we find applying the one-component Logit model on those data coming from the mixture Binomial distribution, cause biased parameter estimations. The MLM can provide much closer estimates than the Logit model, recover the unknown segment membership for the segmentation purpose, and increase the out of sample prediction

TABLE V

CONFUSION MATRIX: CROSS-TABULATION BETWEEN THE HARD SEGMENTATION AND THE CLUSTER FOR THE DATASET *A*

Segment	Estimate	
True	1	2
1	2894	1108
2	66	5932

TABLE VI

CONFUSION MATRIX: CROSS-TABULATION BETWEEN THE HARD SEGMENTATION AND THE CLUSTER FOR THE DATASET *B*

Segment	Estimate	
True	1	2
1	2593	2408
2	199	4800

accuracy. Because the emphasis is on interpretation of segments, we think the MLM for getting unbiased estimates is more valuable than its prediction improvement here. It is worthwhile to check or release the one-distribution assumption of the Logit model, by starting from fitting the MLM.

TABLE VII

THE AVERAGE PROBABILITY IN EACH SEGMENT BY THE CLUSTER FOR THE DATASET *A*

Cluster	The mean probability in seg. 1	The mean probability in seg. 2
1	0.7747112	0.1894139
2	0.2252888	0.8105861

TABLE VIII

THE AVERAGE PROBABILITY IN EACH SEGMENT BY THE CLUSTER FOR THE DATASET *B*

Cluster	The mean probability in seg. 1	The mean probability in seg. 2
1	0.6565434	0.3434566
2	0.3394787	0.6605213

2.5 Dataset and Objective

2.5.1 Dataset Background

The dataset comes from the Direct Marketing Educational Foundation (DMEF), and has 12-year order records from a catalog company, which includes order source, quantity of items, returns, payment information and zip code of the purchaser (zip 5 level). There are 14,448 observations and 44 variables in the original dataset. Because those 14,448 order records repre-

sent 10,000 unique customers, there is not enough records for each customer at his transaction level, we analyze this dataset at order level instead of customer level.

After some exploratory analysis, we decide to utilize this dataset for modeling whether an order is from the print catalog or from the web site.

In real practice, this can help the marketer to forecast future order source distributions between catalog and web, to allocate the limited resources to these two major sources, such as adverting spending, customer services representatives, sales cost and pricing strategy. Further, this type model may assist the marketers to understand the driving factors for a customer using printed catalog or online shopping, to set efficient marketing strategies to meet the customers' wants and needs. For model development and enhancement, we can use this model to compare the Logit model with the MLM. Furthermore, we want to try segment the data based the MLM.

2.5.2 Data Preparation

In addition to catalog orders and web orders, a small amount of the orders are identified as an ancillary item ordered through a telemarketing, direct mail or other promotion. Possible solutions to handle those ancillary orders are to treat them as a separate class in the model prediction, or to merge them with another category. The former method needs to deal with the unbalanced proportion among prediction categories, while the latter should provide a rationale for combining two categories into one. Due to our research focus, the small portion of ancillary orders, and our assumption on the main source of catalog and web in the future, we exclude those ancillary orders from the whole analysis. After that the number of records is reduced to

13,756 from 14,448 for analysis. In other words, 692 (5%) ancillary orders are dropped from our analysis.

In the original dataset, some variables are used only for identification, e.g., 'household ID', and 'order number', so they are excluded from the predictors. Some variables only have one unique value for all the records, such as 'write off amount' and 'payment status code', so these are also excluded. Some variables are highly correlated, such as 'catalog item indicator' and '# of catalog item'; only one of them is kept in the model.

Though Gross Product Revenue (GPR) is the total amount purchased on the order, it contains different components before and after 1/25/2007. Before 1/25/2007, it does not include shipping/handling and taxes. The return and cancel amounts are subtracted from the GPR. Since 1/25/2007 GPR has been changed to include all shipping, handling and tax amounts. Nothing (returns, cancels, etc.) is subtracted from the GPR. To make GPR comparable throughout the whole period, before 1/25/2007, shipping/handling and taxes are added into GPR, and after 1/25/2007 returns and cancels are subtract from GPR.

Some of the variables are utilized to generate other variables, for example, 'input date' (*input_dt*) is used to get 'input year', 'input month' and 'input day'. There are two records with missing values for some fields, we simply delete them without trying some imputations.

After exploring the frequency table of each categorical predictor, we manually exclude '$refund\_rsn_cd$', '$gift\_cert\_redmp\_ind$', '*cupon_redmpind*' and '*add_charg_cd*' from the model building due to extremely unbalanced proportion among possible values. If we allow them in the model building, they will artificially make the model intercept insignificant.

In addition, we delete the orders whose '*payment_cat_cd*' equal '3' or '5', or '*refund_stat_cd*' equal 'C' and 'H'. For those values only have 5, 17, 3 and 1 orders respectively, which will not only cause the similar issue for the intercept, but also the problem for scoring the test data, given there are no such values in the training data.

Finally, the variables we used are 1 target variable and 15 possible predictors. The information about their variable names in our codes, their description, and data type we decided upon is in Table IX. The means and standard deviations of the numerical variables are provided in Table X.

We may treat '*input_yr*' as a categorical variable if we are interested in comparing different year impact on customers purchasing source choice, but we can not apply the model on any orders from future years, because those years are not the possible values in the categorical variable. Treating '*input_yr*' as a numerical variable, we can not only use the model on the future year order based on some assumptions and cautions, but also to some extent see whether there are some trends on customer order source weight change as time goes by. For example, if we see some positive parameter estimate for '*input_yr*' impact on using internet shopping, we may conclude more recent years, the customers are more likely to order through the web channel.

The whole dataset is randomly separated into two parts: Eighty percent of the data are used for model building, and the remaining 20% are reserved for model out of sample prediction performance evaluation.

TABLE IX

THE DATASET VARIABLE DESCRIPTION

Variable Name	Data Description	Type
Target Variable		
item_id	Whether the order is from catalog (1) or from web (0).	Categorical
Predictors		
*add_charg_amt_*	Extra charges for gift wrap or postage.	Numerical
cat_offer	Whether, the type of offer is catalog.	Categorical
gift_ind	Denotes whether or not this order was a gift.	Categorical
*gpr_amt_*	The total amount purchased on this order.	Numerical
input_day	The day of order received.	Numerical
input_mth	The month of order received.	Numerical
input_yr	The year of order received.	Numerical
*nm_of_rct_qty_*	The number of recipients on this order.	Numerical
order_time	The times of orders from same household so far.	Numerical
payment_cat_cd	The payment type.	Categorical
*refund_amt_*	The refund amount for this order.	Numerical
refund_day	The day interval between order received and refund.	Numerical
refund_stat_cd	This code denotes the status of the refund.	Categorical
*sub_qty_*	The quantity of magazine subscriptions purchased on this order.	Numerical
zip_1	The first digit of order zip code.	Categorical

2.5.3 Preliminary Analysis from Statistical Tests

The preliminary analysis we performed consisted of t-tests for numerical predictors between target group 0 and target group 1, and Chi-square test for categorical predictors. Here the t-test is the Welch Two Sample t-test, and the Chi-square test is Pearson's Chi-square test with Yates' continuity correction.

TABLE X

NUMERICAL VARIABLES MEAN AND STANDARD DEVIATION

Variable Name	Mean	Standard Deviation
*add_charg_amt_*	0.8342832	2.282357
*gpr_amt_*	70.20792	58.61592
input_day	14.31899	8.828798
input_mth	9.940608	2.861221
input_yr	2003.724	2.685623
*nm_of_rct_qty_*	1.016647	0.1565689
order_time	1.562518	1.240221
*refund_amt_*	2.199057	13.00254
refund_day	9237.682	2648.478
*sub_qty_*	0.00734225	0.09879626

Although we know there are some limitations to analyze each predictor with regard to the target variable separately, it sometimes can roughly provide some useful information, such as whether all the predictors together can be used to build the model, and which predictors might not be significant in the model. If all the t tests and chi-square tests are insignificant, we will not expect very good model fit from the one-component model. Given that one of numerical or categorical predictors is not different in the test, we should not anticipate it to be a significant predictor in the Logit model.

2.5.3.1 t test for *add_charg_amt_*

Mean in group 0: 0.7071302, Mean in group 1: 0.8963266

$t = -4.1246, d.f. = 7575.399$, p-value = 3.753e-05

Conclusion: Based on the test, we find catalog order extra charge amount is statistical significantly different from web order. Because of the symmetrical of t distribution, and very small p-value, we can even conclude the extra charge of catalog order is statistical significantly higher than web order one. At first, we may think that this makes sense in that usually the cost of handling a catalog order is higher than that for a web order, so to make up the difference or incentivize the customers to use the web source, the company may charge higher or additional fee for those catalog orders. While from the data description, we see those extra charges are for gift wrap or postage, thus what we can get here is the catalog orders have more gifts, assuming all the gift orders will be wrapped.

2.5.3.2 Chi-square test for cat_offer

$\chi - square = 910.7252, df = 1$, p-value$\leq 2.2e-16$

Conclusion: The offer type distributions between catalog order and web order are statistically significant different. By looking at the cross tabulation of cat_offer and $item\_ind$, we find that catalog offer should be the main offer type in the company, and even though the customer places the order via the Internet, he or she still mainly knows the product from the catalog offer. But there are no customers who get the offer through other sources, and place the order by catalog. One reason could be the company sales distribution strategy setting, not allowing people to print out the catalog order form. Another reason could be that once the customers know the offer, the catalog order form is their last way to place order given there are other choices, if that is true, and the company wants to reduce the catalog order volume, in order to sale product more cost-efficiently, we may suggest it increase the weights of other offer sources.

2.5.3.3 Chi-square test for $gift\_ind$

$\chi - square = 35.7166, df = 1$, p-value = 2.282e-09

Conclusion: The $gift\_ind$ distribution between catalog order and web order is statistically significantly different. The catalog orders contain more gift orders. In other words, if the customer want to place a gift order, he or she is more likely to use catalog. This could be due to customer behavior, or the different design between catalog order form and internet order form.

2.5.3.4 t test for $gpr\_amt\_$

Mean in group 0: 67.1873, Mean in group 1: 71.7974

$t = -3.9642, d.f. = 7444.973$, p-value = 7.434e-05

Conclusion: The total amount purchased by catalog is statistically significantly higher than web order. Possible reasons could be: (1) The company charges higher handling fee for catalog orders. (2) The customer using catalog order tends to purchase higher value merchandise.

2.5.3.5 t test for $input\_day$

Mean in group 0: 14.32216, Mean in group 1: 14.33790

$t = -0.088, d.f. = 7223.977$, p-value = 0.9299

Conclusion: There is no evidence to show that order received days between catalog and web are statistically different. To put it another way, there is no trend that early days or later days in the month have more catalog orders or web orders. In addition, we should not see this predictor to be significant in the Logit model.

2.5.3.6 t test for *input_mth*

Mean in group 0: 10.06122, Mean in group 1: 9.91739

$t = 2.4651, d.f. = 6905.046$, p-value = 0.01372

Conclusion: The catalog order receiving month is statistically significantly smaller than the web order. The result shows us that later month in the year, more customers use web to place order. Since p-value here is not very close to 0, we may or may not see '*input_mth*' significant in the model.

2.5.3.7 t test for *input_yr*

Mean in group 0: 2005.126, Mean in group 1: 2003.030

$t = 44.8224, d.f. = 8925.578$, p-value = 2.2e-16

Conclusion: The catalog order receiving year is statistically significantly smaller than the web order. More recent year, more customers utilize web order. The growing popularity of the internet in recent years leads more and more customers to place web order, assuming the company offers similar products, and its customer profile is stable. The company may consider put more investments and resources on its Internet channel.

2.5.3.8 t test for *nm_of_rct_qty_*

Mean in group 0: 1.009695, Mean in group 1: 1.017634

$t = -2.7524, d.f. = 7727.291$, p-value = 0.005929

Conclusion: The number of recipients on catalog order is statistically significantly higher than web order. Tough from the mean value, we see approximately one order for one person in both sources, the more than 1 part can be explained as part of gift receivers, when an order has more

than 1 recipients, it should be a gift order. If that is true, the result still illustrates that people like to use catalog to place gift order, which is similar for $gift\_ind$.

2.5.3.9 t test for $order\_time$

Mean in group 0: 1.485873, Mean in group 1: 1.597260

$t = -4.8548, d.f. = 9500.079$, p-value = 1.224e-06

Conclusion: The times of order for the same household customer from catalog order is statistically significantly higher than web order. It seems that more the customer place an order before, more they tend to stick to catalog order. This may be caused by some kind of the people mental model, once they use some workable method before, and it's still available now, they are hesitate to learn and change their behavior without big enough incentive. We can recommend that when the company wants to promote web order method in the future, they should consider how to offer formal customers big motivation to switch, and treat them differently from those potential customers. Another suggestion will be that the company should look at among those returning customers, what is the proportion of them tried web order before, but switch to catalog order later, if the number is non-ignorable, it should investigate whether there are some flaws in its web order system design.

2.5.3.10 Chi-square test for $payment\_cat\_cd$

$\chi - square = 612.7998, df = 1$, p-value $<$ 2.2e-16

Conclusion: The $payment\_cat\_cd$ value distribution is statistically significantly different between catalog order and web order. Nearly one hundred percent of the customers use a credit card for payment, but there are still some customers who use cash or check to pay the catalog

order. If the customer allow to use both credit card and check (e-check for web) to pay the order, then we find the customer using credit card for order payment is more likely to place order for the web. A little divergence here: even in the recent financial crisis here, people blame the using of credit card, we should not ignore the value of credit card as a payment vehicle, which facilitates many new business developments, such as B to C e-business. Here without the availably of credit card, we may not see many web orders in the company.

2.5.3.11 t test for $refund\_amt$

Mean in group 0: 2.026601, Mean in group 1: 2.234025

$t = -0.7972, d.f. = 7353.527$, p-value = 0.4254

Conclusion: There is no evidence to show refund amount is statistically significantly different between catalog order and web order. This predictor will not be significant in the model. To the customer, it shows both catalog and web descriptions provide the customer similar anticipation for the final product. There is no evidence to indicate that through the web, the customer expects more for the same goods than catalog, but after getting the product, he finds it does not meet his needs or wants.

2.5.3.12 t test for $refund\_day$

Mean in group 0: 9540.979, Mean in group 1: 9101.663

$t = 9.1377, d.f. = 9393.273$, p-value = 2.2e-16

Conclusion: The days between initial order received and final refund are statistically significantly different for catalog order and web order. While we should be cautious to conclude that it takes web order customers longer time to get refund when they decide to return the product.

Maybe it costs a little bit longer time for those customers to get the product, or longer time for them to decide to return the product. If there are some concerns for the company, it can do some survey or focus group study of the web order users, to see whether they are dissatisfied with at the return process, especially the longer time to get a refund.

2.5.3.13 Chi-square test for $refund\_stat\_cd$

$\chi - square = 127.3826, df = 2$, p-value $<$ 2.2e-16

Conclusion: The $refund\_stat\_cd$ distribution is statistically significantly different between catalog order and web order. The difference arises from possible T (written-off) status in the catalog orders, while this value does not appear in the web orders, which could be because of the company policy that only allows T in the catalog orders refund.

2.5.3.14 t test for $sub\_qty\_$

Mean in group 0: 0.0005540166, Mean in group 1: 0.0101736300

$t = -6.8054, d.f. = 8527.634$, p-value = 1.076e-11

Conclusion: The sub_qty of catalog order is statistically significantly higher than web order. From the data dictionary, we know sub_qty means the quantity of magazine subscriptions purchased on the order, and the result is easy to interpret: those customers using web order should be more familiar with the Internet, since there are many electronic magazines in the Internet; they are less likely to subscribe the non-free magazine.

2.5.3.15 Chi-square test for zip_1

$\chi - square = 54.7497, df = 9$, p-value = 1.359e-08

Conclusion: The first digit of zip code distribution between catalog order and web order is

statistically significantly different. If we assuming the company catalog solicitation quantities and website advertisements are proportional to each regions potential customer, we can conclude that the customers from different regions tend to use different sources to place order, because of their different demographical and geographical characteristics.

2.5.4 The Full Logit Model for the Data

We first fit a full Logit model with all the fifteen predictors, by using the glm() package in R. The output is in Figure 9. From the result, we notice some predictors are significant, while some are insignificant. The BIC of the full Logit model is 10892.9, and the hit rate on the reserved test dataset is 73.85%. The cross-tabulation between true value and predicted value is Table XI.

TABLE XI

CROSS-TABULATION BETWEEN TRUE VALUE AND PREDICTED VALUE BY THE FULL LOGIT MODEL

	Prediction	
True Value	0	1
0	395	507
1	211	1633

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5134  -0.9761   0.4631   0.8354   1.8270

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       6.515e+02  1.858e+02   3.507 0.000454 ***
refund_stat_cdT   1.161e+01  2.306e+02   0.050 0.959850
refund_stat_cdZ -5.552e+00  2.075e+01  -0.268 0.789009
gift_indY         8.092e-01  9.161e-02   8.833  < 2e-16 ***
payment_cat_cd2 -5.431e+00  7.092e-01  -7.657 1.90e-14 ***
input_day         3.556e-04  2.678e-03   0.133 0.894365
input_mth        -2.433e-02  8.772e-03  -2.773 0.005548 **
input_yr         -3.310e-01  1.033e-02 -32.031  < 2e-16 ***
sub_qty_          2.544e+00  7.337e-01   3.467 0.000525 ***
nm_of_rct_qty_    2.366e-01  1.924e-01   1.230 0.218765
gpr_amt_          1.269e-03  4.112e-04   3.086 0.002028 **
refund_day        5.562e-04  2.082e-03   0.267 0.789298
refund_amt_       1.556e-03  2.736e-03   0.569 0.569453
cat_offerY        1.744e+01  1.846e+02   0.094 0.924726
add_charg_amt_   -3.997e-03  1.016e-02  -0.393 0.694054
order_time        1.543e-01  2.225e-02   6.933 4.13e-12 ***
zip_11            1.758e-01  8.910e-02   1.973 0.048529 *
zip_12            1.247e-03  9.684e-02   0.013 0.989725
zip_13           -8.285e-03  9.755e-02  -0.085 0.932312
zip_14            3.792e-01  9.679e-02   3.918 8.93e-05 ***
zip_15            1.131e-01  1.228e-01   0.921 0.357027
zip_16            1.433e-01  1.052e-01   1.362 0.173129
zip_17            2.830e-02  1.086e-01   0.261 0.794408
zip_18           -3.300e-01  1.112e-01  -2.967 0.003009 **
zip_19            6.634e-02  9.395e-02   0.706 0.480129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13909  on 10981  degrees of freedom
Residual deviance: 10660  on 10957  degrees of freedom
AIC: 10710
```

Figure 9. The estimates of the full Logit model

2.5.5 The Reduced Logit Model for the Data

Since some of parameter estimations in the full Logit model are insignificant, we want to try whether we can reduce the full Logit model. Here we use backward selection method, and in each selection step BIC is used for variable selection. The output is in Figure 10. Six out of the fifteen variables are retained in the reduced Logit model. The BIC of the reduced Logit model is 10791.85, and the hit rate on the reserved test dataset is 73.53%. The cross-tabulation between true value and predicted value is Table XII.

TABLE XII

CROSS-TABULATION BETWEEN TRUE VALUE AND PREDICTED VALUE BY THE REDUCED LOGIT MODEL

	Prediction	
True Value	0	1
0	379	523
1	204	1640

2.5.6 The Full MLM for the Data

We first fit the MLM with all fifteen predictors, and try the number of components from $K = 2$ to 5. The BIC values of these MLM models are 10933.48, 11089.11, 11279.61 and 11164.8 respectively, which are all larger than the full Logit model BIC. It indicates that the

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5640  -0.9866   0.4666   0.8444   1.6767

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     648.31360  114.15667   5.679 1.35e-08 ***
gift_indY         0.80089    0.09016   8.883  < 2e-16 ***
payment_cat_cd2  -5.68657    0.70870  -8.024 1.02e-15 ***
input_yr         -0.32871    0.01019 -32.262  < 2e-16 ***
sub_qty_          2.60382    0.73263   3.554 0.000379 ***
cat_offerY       16.45995  112.30977   0.147 0.883480
order_time        0.16166    0.02213   7.305 2.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13909  on 10981  degrees of freedom
Residual deviance: 10727  on 10975  degrees of freedom
AIC: 10741
```

Figure 10. The estimates of the reduced Logit model based on BIC backward selection

one-distribution assumption is valid for the full Logit model, and MLM with more than one components is not necessary given the fifteen predictors.

2.5.7 The Reduced MLM for the Data

Here we fit the MLM with those six predictors from the reduced Logit model, and try the number of components from 2 to 5. The BIC of these reduced MLM models are 10732.89, 10785.53, 10844.48 and 10907.48. The reduced two-component MLM has lower BIC than the reduced Logit model does. It has the lowest BIC among all the fitted models. Based on the BIC, the reduced two-component MLM model is our final best model here. The estimates are in Figure 11. The hit rate on the reserved test dataset is 73.49%. The cross-tabulation of true value and predicted value is Table XIII.

TABLE XIII

CROSS-TABULATION BETWEEN TRUE VALUE AND PREDICTED VALUE BY THE REDUCED MLM MODEL

	Prediction	
True Value	0	1
0	376	526
1	202	1642

```
> summary(ff_2_k2)

Call:
stepFlexmix(cbind(item_ind, abs(item_ind - 1)) ~ gift_ind + payment_cat_cd +
    input_yr + sub_qty_ + cat_offer + order_time, data = data8_train,
    control = list(iter.max = 1000, verbose = 100), model = Model1,
    k = 2, nrep = 3)

       prior size post>0 ratio
Comp.1 0.236 2066   8647 0.239
Comp.2 0.764 8916  10980 0.812

'log Lik.' -5296.665 (df=15)
AIC: 10623.33   BIC: 10732.89

> parameters(ff_2_k2)
                            Comp.1      Comp.2
coef.(Intercept)     4988.4894012 938.9061823
coef.gift_indY          3.0358428   1.5090933
coef.payment_cat_cd2  -35.6343249  -4.5433047
coef.input_yr          -2.4817760  -0.4747990
coef.sub_qty_           8.0361672  15.2838782
coef.cat_offerY        10.8530766  18.4035896
coef.order_time         0.8016835   0.2547583
```

Figure 11. The estimates of the reduced MLM

2.5.8 A Segmentation Based on the Reduced MLM

After getting the estimates of the reduced MLM, and following Equation 2.2, we can calculate each case's posterior membership probability, which is the basis of the segmentation in this research. Equation 2.1 provides a way to profile the segment.

Because we do not have typical demographic and geographic variables in the given dataset, nor the capability to collect some information and merge them back to each case, we decide to explore those nine out of the fifteen predictors, which are used in the full Logit model but do not remain in the reduced MLM model. There are two reasons: those nine predictors are not used in the reduced MLM model, and usually the variables for profiling segments should be different for the variables for segmentation; the full Logit model does not need more than one component, while the reduced Logit model does, maybe because some of those nine variables contain some information to describe the hidden segment.

Those nine variables we used are $refund\_stat\_cd$, $input\_day$, $input\_mth$, $nm\_of\_rct\_qty\_$, $gpr\_amt\_$, $refund\_day$, $refund\_amt\_$, $add\_charg\_amt\_$. $refund\_stat\_cd$ and zip_1 are categorical variables, thus we look at the meaning probability in segment 1 within each possible value, and the result is showed in Table XIV and Table XV; the remaining numerical variables are used to calculate their mean within each segment, and the result is report in Table XVI. Comparing to the grand mean of probability in segment 1 0.236, the value '4' and '8' of zip_1 are different. The mean values of $gpr\_amt\_$, $redund\_amt$ and $add\_charg\_amt\_$ differ between segment 1 and segment 2.

Overall, we describe the segments from the reduced MLM model as follows.

(1) In terms of their final channel decision, the customers in segment 1 are more likely to use the web than those from segment 2.

(2) With respect to the predictors for making a decision, the customers in segment 1 are more sensitive to whether it is a gift, payment type, and the order time in their purchase history; while the buyers in segment 2 are more likely to be affected by the quantity of magazine subscriptions and whether received the catalog offer.

(3) To profile the segments, here we can define the customers in segment 2 as high spenders, because they have higher total purchase and additional charge, accordingly higher refund, and they are more likely from west states whose first zip code is '8'. Accordingly, the customers in segment 1 can be called as low spenders.

It is worth mentioning another segmentation method based on the Logit model in common practice. At first a Logit model is built, then the model is used to score each case to get the response probabilities, and assign each case based on the model score, such as into ten deciles or use 0.5 as a cut-off point, finally profiling is done for each segment. This method seems similar to our MLM based method, in that both utilize the Logit model, and segmentation is in relation to a choice result, while our method is different from current real practice method in three important aspects. The first one is that we use a model selection criterion to decide the number of components or segments, the second one is that the between-segment difference mainly shows in the parameters difference, instead of the different success rates with respect to the dependent variable, and the third one is that the MLM based segmentation does segmentation and Logit

model fitting simultaneously. We think our MLM based segmentation is more objective than current practice segmentation.

TABLE XIV

THE AVERAGE PROBABILITY IN SEGMENT 1 WITH $REFUND\_STAT\_CD$

Value	Average segment 1 probability
P	0.2288020
T	0.2358893
Z	0.2362261

2.6 Summary and Future Research

In this chapter, we started from the regular Logit model and the FMM, and then introduced the MLM. We think the one-Binomial distribution assumption of the Logit model needs to be checked, and the MLM allows the data from the mixture distribution of more than one Binomial distribution. Furthermore, we think the MLM can be used to generate segments on the basis of from which distribution the data come. Although we used the existing FlexMix package in R for the MLM fitting, we provide an E-M algorithm to estimate the MLM. Two datasets were simulated from mixture Binomial distributions with the success rate being modeled through the Logit link function. The Logit model and the MLM model were utilized to fit the data. The results show that the MLM model provides unbiased estimates, while the Logit model gives

TABLE XV

THE AVERAGE PROBABILITY IN SEGMENT 1 WITH ZIP_1

Value	Average segment 1 probability
0	0.2419564
1	0.2249501
2	0.2454876
3	0.2450782
4	0.2078296
5	0.2312233
6	0.2295646
7	0.2389392
8	0.2803469
9	0.2360563

biased estimates. There are only slight differences between the Logit model and the MLM out-of-sample prediction performance. The MLM does recover the true distribution membership more accurately than random guessing.

We further studied a dataset from the DMEF, which contains some order information. We fitted models to predict whether a customer places an order through the web or the catalog. The goal is to see which variables predict this, and whether their coefficients may differ across segments. After some necessary data preparation, and preliminary analysis based on the statistical test, we built a full Logit model, a reduced Logit model, a full MLM and a reduced MLM. All of those model have the similar out-sample prediction performances, but the reduced MLM model has the lowest BIC, which is our final model. We segmented the data on the basis of the

TABLE XVI

SEGMENT 1 AND SEGMENT 2 AVERAGE VALUE OF SOME VARIABLES

Variable	Mean in segment 1	Mean in segment 2
item_ind	0.2390743	0.8046547
input_day	14.31650	14.33773
input_mth	10.04987	9.938377
*nm_of_rct_qty_*	1.012471	1.015813
*gpr_amt_*	68.79235	70.74165
refund_day	9261.911	9241.188
*refund_amt_*	2.019428	2.211022
*add_charg_amt_*	0.7602088	0.8569466

reduced MLM, and got two segments. Further, we describe these segments from the order channel decision, the sensitivity to the predictors, and the difference of other extra variables which are not used in the reduced MLM. We also think our segmentation method is more subjective and systematically, comparing to some real practice segmentation from the Logit model. In summary, in this research we suggested that the one-distribution assumption of the Logit model be checked by trying the MLM, and proposed a different MLM-based segmentation method.

In future research, we can try to use different methods to search for the optimal value of the MLM likelihood function, and compare the results. In addition to BIC, different model selection criterion can also be used for deciding the number of components. The Logistic Regression Tree (LRT) model can be of interest to study and compare with the MLM. Other marketing datasets which contains different segments in theory, but no explicit given variables for the segments

are worthwhile to study, such as the response to the cross-sell product in the financial services industry.

CHAPTER 3

A SEGMENTATION OF S&P 500 MONTHLY RATES OF RETURN USING THE HIDDEN MARKOV MODEL

3.1 Introduction

We study monthly S&P 500 index data from January 1950 to August 2010. This dataset includes open, index, index, low, and adjusted close.

We are interested in the monthly rates of return (ROR), which could be modeled by the Hidden Markov Model (HMM). More importantly, through the hidden states of the HMMs, we may find the market changes among different hidden states from time to time, and define those hidden states as some kind of market segment or phase.

The continuous ROR is defined as:

$$ROR_t = \ln P_t - \ln P_{t-1}, \tag{3.1}$$

where

t indexes the months

P_t is adjusted close from month t

The series $\ln P_t$ is found to be non-stationary by the Phillips-Perron Unit Root test, but its first order difference (continuous ROR) is a stationary series based on the same test.

The continuous ROR Equation 3.1 approximates (from below) the discrete one, which is defined as

$$\text{discrete } ROR_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{3.2}$$

Let $x = P_t/P_{t-1}$ and recall that by Taylor series, $f(x) \sim f(a) + (x-a)f'(a)$. When x is close to a, take $f(x) = \ln x$ and $a = 1$, then $\ln x \sim x - 1$, write

$$\begin{aligned}
\text{discrete } ROR_t &= \frac{P_t - P_{t-1}}{P_{t-1}} \\
&= \frac{P_t}{P_{t-1}} - 1 \\
&= x - 1 \\
&\sim \ln x \\
&= \ln \frac{P_t}{P_{t-1}} \\
&= \ln P_t - \ln P_{t-1} \\
&= \text{continuous } ROR_t
\end{aligned}$$

Monthly data are used because the segmentation obtained may be used in future research relating to the financial analyst's beta, which is estimated from monthly data. (The customary way to estimate a stock's "beta" is to use five years of monthly data).

Maheu and McCurdy (2000) used a Markov-switching model that incorporates duration dependence to capture nonlinear structure in both the conditional mean and the conditional variance of stock returns. The model sorts returns into a high-return stable state and a low-

return volatile state. These correspond to the usual Bull and Bear markets, respectively and are labeled as such. The monthly returns include dividends and range from 1802 to 1995 (33).

Lunde and Timmermann (2004) studied time series dependence in the direction of stock prices by modeling the (instantaneous) probability that a Bull or Bear market terminates as a function of its age and a set of underlying state variables, such as interest rates. The strongest effect of increasing interest rates is found to be a lower Bear market hazard rate and and hence a higher likelihood of continued declines in stock prices (31).

3.2 Three Plots for Time Series Analysis

First we want to check whether an ARIMA model can be used to fit the ROR data. The time series plot of monthly ROR is Figure 12. The ACF plot of the monthly ROR data is showed in Figure 13, and the PACF plot of the monthly ROR data is given in Figure 14.

The ACF and PACF of monthly ROR show no sizeable lag correlations; therefore it is inappropriate to model the ROR data by using an ARIMA model. (If it had been appropriate to fit an ARIMA model, we could have benchmarked the HMM model against the ARIMA model.)

3.3 A Simple Linear Time Trend Regression

After we conclude that there are no auto-correlations among monthly ROR, we can try to fit a simple regression between monthly ROR and increasing time t ($t = 1, ..., 728$) , so that we can check whether there is a time trend in the data, or whether a simple linear time trend regression is good enough to model the data.

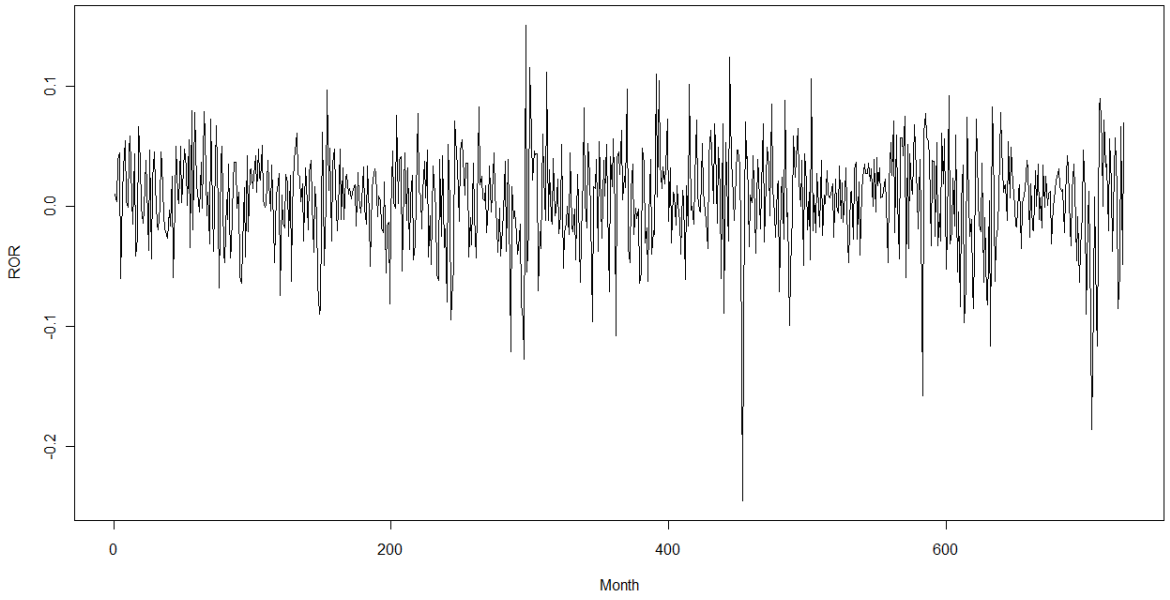

Figure 12. Time Series Plot of the Monthly *ROR* from February 1950 to August 2010

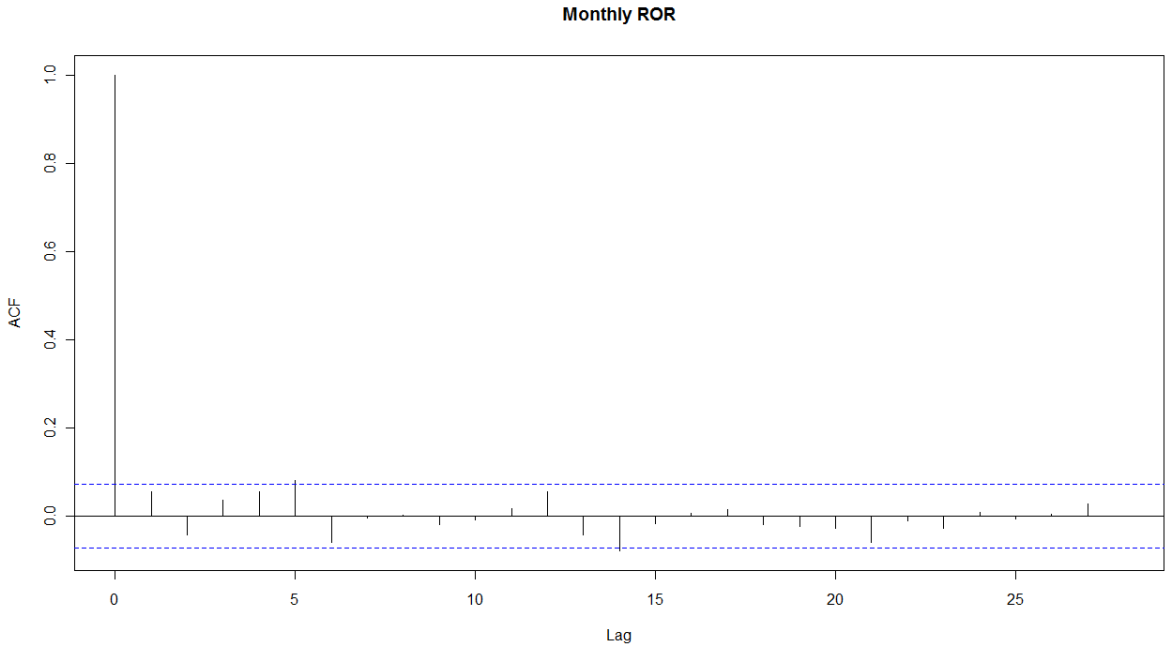

Figure 13. The ACF Plot of the Monthly *ROR* Data

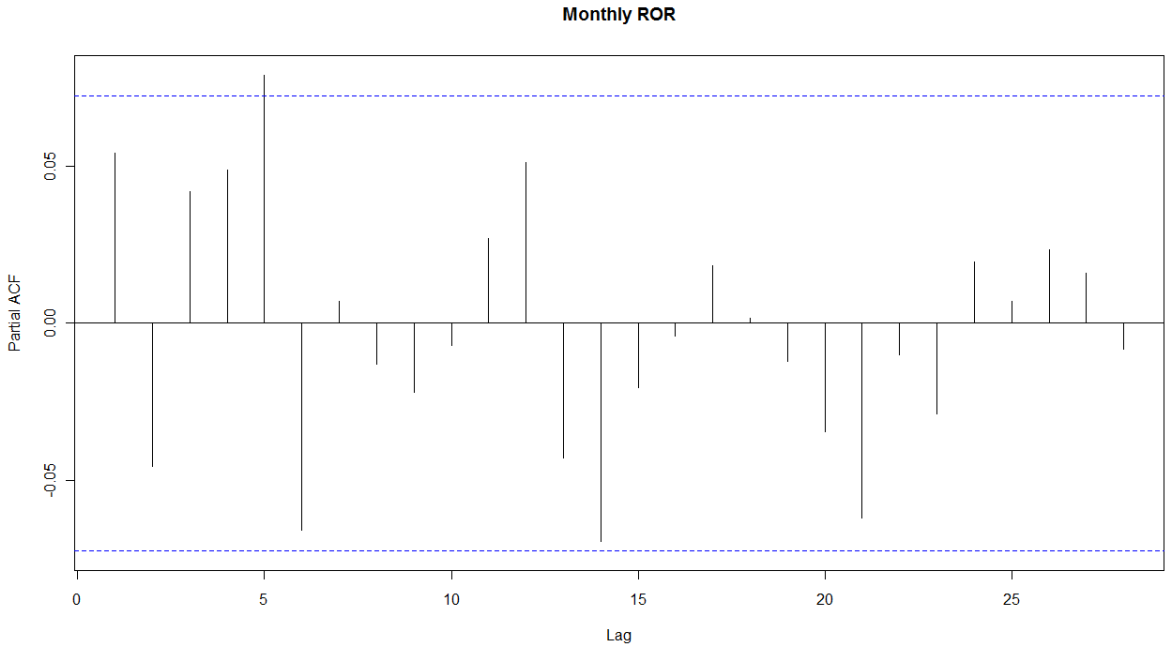

Figure 14. The PACF Plot of the Monthly *ROR* Data

```
Call:
lm(formula = x ~ t)

Residuals:
      Min        1Q    Median        3Q       Max
-0.250599 -0.023601  0.003267  0.029100  0.144825

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.032e-03  3.141e-03   2.557   0.0108 *
t           -6.255e-06  7.466e-06  -0.838   0.4024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04233 on 726 degrees of freedom
Multiple R-squared: 0.0009658,  Adjusted R-squared: -0.0004103
F-statistic: 0.7019 on 1 and 726 DF,  p-value: 0.4024
```

Figure 15. The OLS Output of lm() from R

We set the monthly ROR as the target variable and t as the explanatory variable, then use the 'lm( )' procedure in R to fit the linear regression. The output from R is in Figure 15.

The insignificance of the parameter estimate for regression on t indicates that there is no time-trend effect in the data. The tiny value of R^2 (0.0009658) and big p-value (0.4024) of the F-test illustrate the simple linear time trend model is insufficient to fit the data.

To visualize the fit of the linear regression, we plot the regression curve within the data dot plot, and list it in Figure 16.

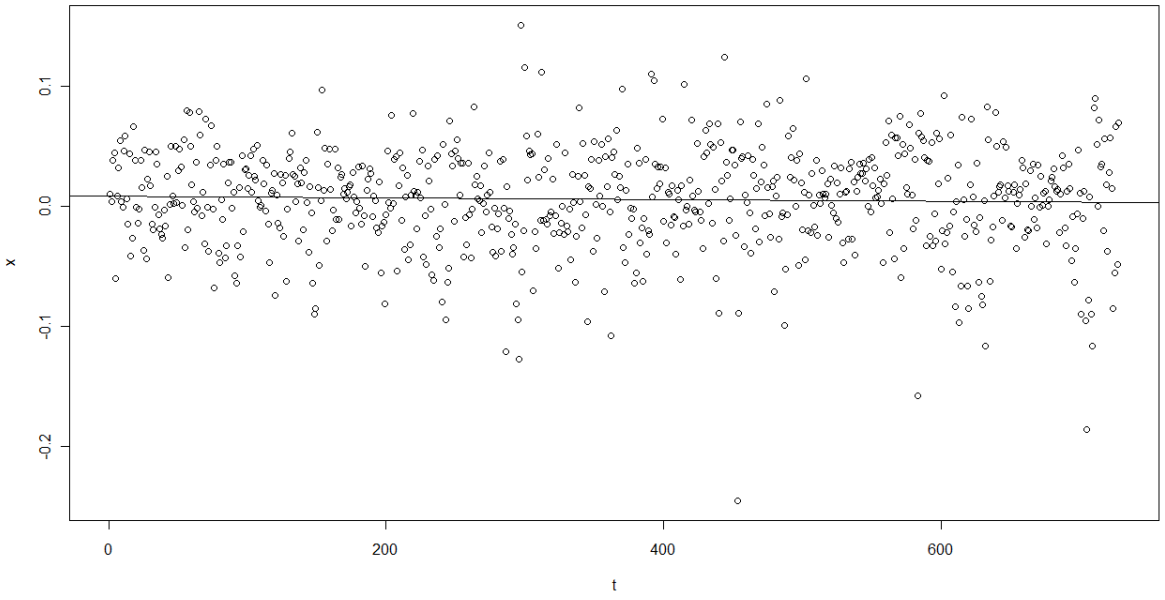

Figure 16. The OLS Fitted Curve within Data Points

3.4 A Segmentation Based on the Hidden Markov Model

Through the hidden states of the fitted HMM, we can obtain a model-based market segmentation. In other words, we let the data suggest the properties of Bull and Bear states, see if the parameter estimates correspond to conventional ideas of Bull and Bear, and let the HMM label each time point as Bull or Bear. In this research, we assume that the state dependent distribution is Normal, and will try from 1 to 3 states. Considering only 1 hidden state is equivalent to letting the monthly ROR follow the Normal distribution, thus we need to check the data Normality.

Before the HMM is fitted, we expect the characteristics differences among different hidden states should be one of these three possible scenarios: (1) the state dependent Normal distributions means μ_k are different, (2) the state dependent Normal distributions standard deviations σ_k are different, (3) both μ_k and σ_k are different.

3.4.1 Elements of an HMM

An HMM is characterized by the following:

(1) K, the number of hidden states in the model.

(2) T, the length of observation sequence.

(3) $\pi_k = \Pr(S_1 = k)$, the probability of being in state k at the beginning i.e. at $t = 1$.

(4) $P = \{p_{ij}\}$, the state transition probability matrix, where $p_{ij} = \Pr(S_{t+1} = j | S_t = i)$, the probability of be in state j at time $t + 1$ given that we were in state i at time t.

(5) $F = \{f_k\}$, where f_k is the probability density function or the probability mass function give

$S = k$.

(6) Y_t, the observed outcome at time t.

Rabiner (1989) is perhaps the first major published comprehensive work on HMMs (41). The technical report by Dugad (1996) is a tutorial presenting an overview of what are HMMs, what are the different problems associated with HMMs, the Viterbi algorithm for determining the optimal state sequence, algorithms associated with training HMMs, and distance between HMMs (15).

The HMM is an extension of the FMM because the observed y can be considered from a mixture distribution in the following way:

$$f(y_t) = p_t(1)f_1(y_t) + p_t(2)f_2(y_t) + \cdots + p_t(k)f_k(y_t) + \cdots + p_t(K)f_K(y_t)$$

where $p_t(k) = \Pr(S_t = k)$, and

$$f(y_t|S_t = k) = p_{1k}f_1(y_t) + p_{2k}f_2(y_t) + \cdots + p_{kk}f_k(y_t) + \cdots + p_{Kk}f_K(y_t)$$

where $p_{ik} = \Pr(S_t = k|S_{t-1} = i)$

We consider the HMM is one kind of dynamic FMM, because it allows transition among segments.

3.4.2 One-state HMM

Assuming the state dependent distribution is Normal, the estimation of a one-state HMM is easy in that we only need to get the μ and σ through maximum likelihood estimation and

use the well-known Equation 3.3 and Equation 3.4. Given the data, here $\hat{\mu}$ is 0.00575261 and $\hat{\sigma}$ is 0.04232393.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.3}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2 \tag{3.4}$$

Whether the data follow a Normal distribution is a critical assumption which we need to check. Next we will use different available Normality tests and q-q plot in software R to check this assumption. There are six Normality tests in the 'nortest' package of the R software, which are the Anderson-Darling test, Cramer-von Mises test, Lilliefors (Kolmogorov-Smirnov) test, Pearson chi-square test, Shapiro-Francia test and Shapiro-Wilk. The null hypothesis of all those tests is that the data are Normally distributed, and the results are in Table XVII. The Q-Q plot is showed in Figure 17.

TABLE XVII

THE NORMALITY TESTS TABLE.

Test Name	Test Statistic	P-Value
Anderson-Darling normality test	A = 3.0201	1.380e-07
Cramer-von Mises normality test	W = 0.4807	4.153e-06
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.0457	0.001043
Pearson chi-square normality test	P = 45	0.008362
Shapiro-Francia normality test	W = 0.9714	1.033e-09
Shapiro-Wilk normality test	W = 0.9734	3.016e-10

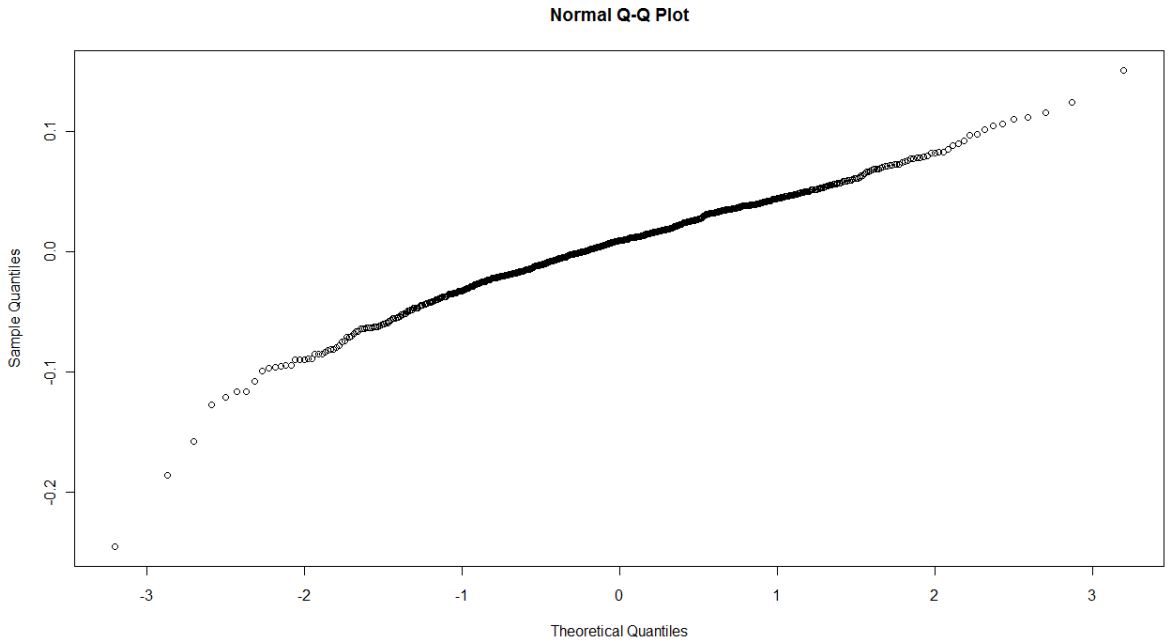

Figure 17. The Normal Q-Q plot of monthly *ROR*

Even though these Normality tests have different formulas to calculate the test statistics, or their own assumptions to derive the p values, all the p values are less than 0.05. Looking at the Q-Q plot, it clearly shows that the curve is not straight. Based on the above results, we consider the data are not Normally distributed, and a one-state HMM is obviously not appropriate for the given monthly *ROR* data. It still can be the case that the class-conditional distributions are Normal for more than one class.

3.4.3 The Estimation of the HMM

Following the method in the book by Zucchini and McDonald, we use Maximum Likelihood Estimation (MLE) method to estimate the HMM parameters in this research. There are three major problems to solve by using MLE method. The first one is what the likelihood function is, the second one is how to maximize the likelihood function, and the third one is how to prevent the numerical underflow or overflow issue.

3.4.3.1 The Likelihood Function of the HMM

The likelihood function L means the probability of the given data have been observed under the HMM parameters, and the formula is shown in Equation 3.5.

$$P(X|\pi, P, F) = \pi F_1 \sum_{t=2}^{T} P_t F_t, \tag{3.5}$$

where

π is a 1 by K initial probability vector

F_t is a K by K state-dependent density matrix

P_t is a K by K transition probability matrix

3.4.3.2 Three Ways to Maximize the HMM Likelihood Function

Because the objective function includes higher order terms of the free parameters, and the form is non-linear, there is no explicit analytical solution; we can only find some kind of numerical solution. The problem is one of optimization of a nonlinear function.

There are three ways to find a numerical solution for the HMM MLE result according to the current research. The first one is to use a Newton type algorithm to search a solution. The second one is to use Expectation-Maximization (EM) algorithm to find a solution. The third one is to utilize Markov Chain Monte Carlo (MCMC) simulation to get a solution.

Both EM and MCMC require the derivations from the statistical knowledge, but for EM, sometime there is still no analytic form for the interim result in the Maximization step, the numerical optimization is still required. The convergence speed of EM is slower than Newton type algorithm, and it still cannot guarantee the global optimum is found. When applying MCMC method, some prior distributions with the given parameters have to set for the free parameters, but which prior distributions are the proper ones are difficult to choose objectively. A warning note regarding MCMC has been given by Celeux, Hurn and Robert (2000) (5), and repeated by Chopin (2007) (10): we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge! Although this statement was made for mixture models, it causes some concern about using the MCMC method for the HMM estimation, because the HMM is one type of finite mixture model. When we use the Newton type algorithm to search a possible solution, we avoid the above issues, and also give us the room to build more complicated HMM, as long as the log-likelihood function has the explicit form and the underflow or overflow issue is prevented.

3.4.3.3 A Scaling Technique to Prevent Overflow or Underflow Issue

Another challenge for HMM estimation is to prevent overflow or underflow problem as the length of sequence T gets large. The overflow problem occurs in these continues state-dependent

distributions, while the underflow issue happens in those discrete state-dependent distributions. The classical method is to take the log of the likelihood function, which is insufficient according to the HMM literature. In addition, some kind of scaling technique has to be applied when calculating the cumulative log likelihood at each time point t. Recalling that the forward probability $\alpha_1 = \pi F_1$ and $\alpha_t = \alpha_{t-1} P_t F_t$ for $t > 2$. The idea here is to continuously scale α to prevent overflow or underflow. The algorithm for writing the scaled log-likelihood function is obtained from the former literature review, and can be illustrated below:

When $t = 1$:

Step 1: Calculate α_1, a 1 by K vector;

Step 2: Divide each α_1 element by the sum of α_1 elements, to get α_1';

Step 3: Let the log-likelihood L_1 be the log of the sum of α_1 elements.

When $t > 1$:

Step 1: Calculate $\alpha_t = \alpha_{t-1}' P_t F_t$. Notice that we use the scaled α_{t-1}' instead of the original α_{t-1} in the calculation;

Step 2: Divide each α_t element by the sum of α_t elements, to get α_t';

Step 3: Let the log-likelihood L_t is the log of the sum of α_t elements, and plus L_{t-1}. Notice that here we still use α_t instead of the scaled α_t'.

Repeat the above steps until $t = T$.

This kind of scaled log-likelihood function is scaled through α, in the way that the current time point log-likelihood function is scaled by previous time point scaled α_{t-1}', or the current time point scaled α_t' is utilized for next time point log-likelihood function calculation.

3.4.3.4 Other Problems in Estimating the HMM

After choosing the Newton type algorithm to maximize the log-likelihood function, we still face two problems. One problem is how to get the global optimum instead of a local optimum, and the other is how to deal with the constrained optimization problem.

The first problem is a notoriously difficult problem for numerical optimization, and its difficulty depends on the shape of objective function. What we can do here is to improve the possibility of reaching global optimum. The suggestion we adopt is to try sets of different initial values, and take the solution with best value for the objective function.

The second problem appears because some of the HMM parameters have to be within the 0 to 1 range, such as the initial probability π. The constrained optimization can be solved in two ways. One is to use some optimizers which can set the range for those free parameters; the other is to create some unbounded working parameters when optimizing the function, and then make some transformations of those working parameters to get the original free parameters. When using the specific optimizers for constrained optimization, sometime it is hard to find a solution given there are some of the parameters close to their boundaries. Therefore, in this research we create some working parameters during the optimization process, and then transform them back to the original parameters.

3.4.4 Two-state HMM

Once we assume there are two hidden states in the HMM, and each state-dependent distribution is Normal, then there are seven free parameters in the model. They are the following:
(1) The initial probability in the state 1 π_1;

(2) The staying probability of the state 1 p_{11};

(3) The staying probability of state 2 p_{22};

(4) The mean of the state 1 normal distribution μ_1;

(5) The standard deviation of the state 1 normal distribution σ_1;

(6) The mean of the state 2 normal distribution μ_2;

(7) The standard deviation of the state 2 normal distribution σ_2.

The reaming three parameters are derived in the following way:

(1) The initial probability in the state 2 $\pi_2 = 1 - \pi_1$;

(2) The transition probability from the state 1 to the state 2 $p_{12} = 1 - p_{11}$;

(3) The transition probability from the state 2 to the state 1 $p_{21} = 1 - p_{22}$.

The likelihood function is in Equation 3.6.

$$P(X|\pi, P, F) = \pi F_1 \sum_{t=2}^{T} P_t F_t, \tag{3.6}$$

where

$\pi = (\pi_1, \pi_2)$

$P_t = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$

$F_t = \begin{pmatrix} f_1 & 0 \\ 0 & f_2 \end{pmatrix}$

$f_k = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$

To reduce the unnecessary computation, we remove the constant term $\frac{1}{\sqrt{2\pi}}$ from the normal density function when we write the code in the R software.

Five out of seven free parameters are bounded in the two-state HMM, and we create five working parameters, and make the following transformations to the free parameters. The details are as follows:

(1) Since π_1 is between 0 and 1, so let $\pi_1 = \frac{\exp(\alpha_\pi)}{1+\exp(\alpha_\pi)}$

(2) Since p_{11} is between 0 and 1, so let $p_{11} = \frac{\exp(\alpha_{p_{11}})}{1+\exp(\alpha_{p_{11}})}$

(3) Since p_{22} is between 0 and 1, so let $p_{22} = \frac{\exp(\alpha_{p_{22}})}{1+\exp(\alpha_{p_{22}})}$

(4) Since σ_1 is greater than 0, so let $\sigma_1 = \alpha_{\sigma_1}^2$

(5)Since σ_2 is greater than 0, so let $\sigma_2 = \alpha_{\sigma_2}^2$

Recall that we expect there might be some difference between μ_1 and μ_2, plus in order to prevent the label switching problem, we therefore predefine $\mu_1 \leq \mu_2$, and let $\mu_2 = \mu_1 + \exp(\alpha_\mu)$ in our code.

We use the 'nlm' function in the R package to minimize the negative log-likelihood function, run ten different initial values, and finally get a solution with best objective function value. The estimates of the two-state HMM parameters are in the followings:

$$\hat{\pi} = (0.0000014, 0.9999986)$$

$$\hat{P} = \begin{pmatrix} 0.85419957 & 0.1458004 \\ 0.03386228 & 0.9661377 \end{pmatrix}$$

$$\hat{\mu}_1 = -0.01418582$$

$$\hat{\sigma}_1 = 0.06460005$$

$\hat{\mu}_2 = 0.01025555$

$\hat{\sigma}_2 = 0.03376595$

The scaled log-likelihood is 1976.742. Since the length of the sequence T is 728, and the number of free parameter p is 7, so the BIC is calculated in Equation 3.7.

$$\begin{aligned} BIC &= -2\ln L + p\ln n \\ &= -2 \times 1976.742 + 7 \times \ln 728 \\ &= -3907.352 \end{aligned} \tag{3.7}$$

Before we analyze the parameter estimates from the two-state HMM, let us proceed to the three-state HMM, then choose the suitable model based on the BIC.

3.4.5 Three-state HMM

Assuming there are three hidden states in the HMM, then there are fourteen free parameters. Actually there are $k-1$ free parameters for the initial probability vector, $k \times (k-1)$ free parameters for the transition probability matrix, and $2 \times k$ free parameters for the state-dependent normal distributions. Therefore, the number of parameters is fourteen Equation 3.8.

$$\begin{aligned} p &= (k-1) + k \times (k-1) + 2 \times k \\ &= k^2 + 2k - 1 \end{aligned} \tag{3.8}$$

Those fourteen parameters we need to estimate are the followings:

(1) The initial probability in the state 1 π_1;

(2) The initial probability in the state 2 π_2;

(3) The transition probability from the state 1 to the state 2 p_{12};

(4) The transition probability from the state 1 to the state 3 p_{13};

(5) The transition probability from the state 2 to the state 1 p_{21};

(6) The transition probability from the state 2 to the state 3 p_{23};

(7) The transition probability from the state 3 to the state 1 p_{31};

(8) The transition probability from the state 3 to the state 2 p_{32};

(9) The mean of the state 1 normal distribution μ_1;

(10) The standard deviation of the state 1 normal distribution σ_1;

(11) The mean of the state 2 normal distribution μ_2;

(12) The standard deviation of the state 2 normal distribution σ_2;

(13) The mean of the state 3 normal distribution μ_3;

(14) The standard deviation of the state 3 normal distribution σ_3.

The remaining four parameters are derived by the following formula:

(1) The initial probability in the state 3 $\pi_3 = 1 - \pi_1 - \pi_2$;

(2) The staying probability of the state 1 $p_{11} = 1 - p_{12} - p_{13}$;

(3) The staying probability of the state 2 $p_{22} = 1 - p_{21} - p_{23}$;

(4) The staying probability of the state 3 $p_{33} = 1 - p_{31} - p_{32}$.

The likelihood function is in Equation 4.1.

$$P(X|\pi, P, F) = \pi F_1 \sum_{t=2}^{T} P_t F_t \tag{3.9}$$

Where

$\pi = (\pi_1, \pi_2, \pi_3)$

$$P_t = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

$$F_t = \begin{pmatrix} f_1 & 0 & 0 \\ 0 & f_2 & 0 \\ 0 & 0 & f_3 \end{pmatrix}$$

$f_k = \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$

To reduce the unnecessary computation, we remove the constant term $\frac{1}{\sqrt{2\pi}}$ from the normal density function when we write the code in the R software.

Eleven out of fourteen free parameters are bounded in the three-state HMM, and we create eleven working parameters, and make the following transformations to the free parameters. The details are in the followings:

(1) Since π_1 is between 0 and 1, so let $\pi_1 = \frac{\exp(\alpha_{\pi_1})}{1+\exp(\alpha_{\pi_1})+\exp(\alpha_{\pi_2})}$;

(2) Since π_2 is between 0 and 1, so let $\pi_2 = \frac{\exp(\alpha_{\pi_2})}{1+\exp(\alpha_{\pi_1})+\exp(\alpha_{\pi_2})}$;

(3) Since p_{12} is between 0 and 1, so let $p_{12} = \frac{\exp(\alpha_{p_{12}})}{1+\exp(\alpha_{p_{12}})+\exp(\alpha_{p_{13}})}$;

(4) Since p_{13} is between 0 and 1, so let $p_{13} = \frac{\exp(\alpha_{p_{13}})}{1+\exp(\alpha_{p_{12}})+\exp(\alpha_{p_{13}})}$;

(5) Since p_{21} is between 0 and 1, so let $p_{21} = \frac{\exp(\alpha_{p_{21}})}{1+\exp(\alpha_{p_{21}})+\exp(\alpha_{p_{23}})}$;

(6) Since p_{23} is between 0 and 1, so let $p_{23} = \frac{\exp(\alpha_{p_{23}})}{1+\exp(\alpha_{p_{21}})+\exp(\alpha_{p_{23}})}$;

(7) Since p_{31} is between 0 and 1, so let $p_{31} = \frac{\exp(\alpha_{p_{31}})}{1+\exp(\alpha_{p_{31}})+\exp(\alpha_{p_{32}})}$;

(8) Since p_{32} is between 0 and 1, so let $p_{32} = \frac{\exp(\alpha_{p_{32}})}{1+\exp(\alpha_{p_{31}})+\exp(\alpha_{p_{32}})}$;

(9) Since σ_1 is greater than 0, so let $\sigma_1 = \alpha_{\sigma_1}^2$;

(10) Since σ_2 is greater than 0, so let $\sigma_2 = \alpha_{\sigma_2}^2$;

(11) Since σ_3 is greater than 0, so let $\sigma_3 = \alpha_{\sigma_3}^2$.

We still think there might be some differences among μ_1, μ_2, and μ_3, so in order to prevent the label switching problem, we predefine $\mu_1 \leq \mu_2 \leq \mu_3$, and let $\mu_2 = \mu_1 + \exp(\alpha_{\mu_1})$ and $\mu_3 = \mu_1 + \exp(\alpha_{\mu_1}) + \exp(\alpha_{\mu_2})$ in our code.

After trying six different initial values, and selecting the one solution provides the smallest objective function, and transferring those working parameters to the original free parameter, then deriving the remaining parameters, we get the following parameter estimates:

$\hat{\pi} = (0, 1, 0)$, i.e., $\hat{\pi}_2 \geq 0.99999999$

$$\hat{P} = \begin{pmatrix} 0.814371 & 0 & 0.1856291 \\ 0.02431853 & 0.9756815 & 0 \\ 0 & 0.5604675 & 0.4395325 \end{pmatrix}$$

$\hat{\mu}_1 = -0.03356817$

$\hat{\sigma}_1 = 0.06621817$

$\hat{\mu}_2 = 0.008703582$

$\hat{\sigma}_2 = 0.03457888$

$\hat{\mu}_3 = 0.05640828$

$\hat{\sigma}_3 = 0.01877439$

The scaled log-likelihood function value is 1987.052. Since the length of the sequence T is 728, and the number of free parameter is 14, so the BIC is calculated in Equation 3.10.

$$\begin{aligned} BIC &= -2\ln L + p\ln n \\ &= -2 \times 1987.052 + 7 \times \ln 728 \\ &= -3881.84 \end{aligned} \tag{3.10}$$

3.4.6 Choice between Two-state and Three-state HMMs

Before we make the selection between two-state and three-state HMMs, we notice that both model results show the bigger μ has the smaller σ.

When looking at the three-state HMM estimates for the transaction probability matrix P, $\hat{p}_{12}$, $\hat{p}_{23}$ and $\hat{p}_{31}$ are all equal to 0. There are no possibilities from the state 1 to the state 2, or the state 2 to the state 3, or the state 3 to the state 1 within one step. In addition $\hat{p}_{33}$ is even less than 0.5, which means the state 3 is an unstable state when time goes by. The three-state model is hard to explain according to domain knowledge, even though there are explicit differences among the means μ and the standard deviations σ.

Finally when referring to the BIC standard, the two-state HMM's BIC (-3907.352) is smaller than the three-state HMM's one (-3881.84). The scaled log-likelihood function value of the three-state HMM (1987.052) is larger than the two-state HMM's one (1976.742), but it needs

more free parameters (14 vs. 7). The trade-off between model's goodness of fit and model's complexity shows the increase of the scaled log-likelihood is not worth fitting more complex model. Certainly people can argue that maybe the estimates of the three-state HMM is not a good local optimum, using another good optimizer could improve the scaled log-likelihood of the three-state HMM, but the estimates of the two-state HMM are from the same optimizer, and using the same different initial values strategy, using another optimizer may also increase its scaled log-likelihood value. How to generate a good optimization method to improve the optimal value of the objective function, is beyond the scope of this research, thus we leave it as a future research direction for our HMM improvement.

In summary, based on both the explanation of the parameter estimates and the BIC criterion, plus assuming both the two-state HMM and the three-state HMM estimates from the objective function local optimal are acceptable now, we select the two-state HMM as our final model.

3.4.7 The Explanation of the Two-state HMM

Recall the estimates of the two-state HMM are the following:

$$\hat{\pi} = (0.0000014, 0.9999986)$$

$$\hat{P} = \begin{pmatrix} 0.85419957 & 0.1458004 \\ 0.03386228 & 0.9661377 \end{pmatrix}$$

$$\hat{\mu}_1 = -0.01418582$$

$$\hat{\sigma}_1 = 0.06460005$$

$$\hat{\mu}_2 = 0.01025555$$

$$\hat{\sigma}_2 = 0.03376595$$

Note in parameter that $\hat{\mu}_1$ is negative, and $\hat{\mu}_2$ is positive, we define the hidden state 1 as the Bear state of the market, and the hidden state 2 as the Bull state. We explain the estimates as below:

(1) Most likely the S&P 500 index from January 1950 starts in the Bull market, since $\hat{\pi}_2 = 0.9999986$.

(2) The market tends to stay in its current hidden state, because $\hat{p}_{11} = 0.85419957$and $\hat{p}_{22} = 0.9661377$. It is more likely to stay in the Bull market than to remain in the Bear market since $\hat{p}_{22} > \hat{p}_{11}$. In other words, recovering from the Bear market to the Bull market is more likely.

(3) Based on the t.p.m. estimation and the Markov chain theory, we can calculate the limiting distribution of those two states P_L by solving the system of linear equations: $P_L = P_L \times P$ and $P_L(1) + P_L(2) = 1$. Since $\hat{P_L} = (0.1884769, 0.811523)$, we expect in the long run, nineteen percent of chance the Market is in the Bear market, and eighty-one percent of possibility is in the Bull market. The expected return time of the state 1, which is the expected number of steps until the chain revisits state 1, is 5.3 months. And expected return time of the state 2 is 1.2 months. The average months of staying in the Bear market is 6.6 months, and the expected months to remain in the Bull market is 29 months.

(4) The average ROR in the Bear market is lower than the one in the Bull market, because $\mu_1 < \mu_2$. Actually the expected mean of ROR in the Bear market is negative, while the mean of ROR in the Bull market is positive. Both signs of our HMM based Bear and Bull market are consistent with most of other definitions of the Bear and the Bull market.

(5) In addition to the majority of current literature, we find the volatility of ROR in the Bear

market is higher than the one in the Bull market, since $\hat{\sigma}_1 > \hat{\sigma}_2$.

(6) From the μ and σ comparisons, it suggests that on the monthly basis it is not a good idea to invest in the Bear market, because its expected value is negative, and its risk is high.

3.4.8 Hidden States Recovery at Each Time Point

After we get the two-state HMM parameters estimates, we can utilize them to calculate the state probability by using Equation 3.11:

$$\begin{aligned} P(S_t = k|\pi, P, F, X) &= \frac{\alpha_t(k)\beta_t(k)}{\sum_{k=1}^{K} \alpha_t(k)\beta_t(k)} \\ &= \frac{\alpha_t(k)\beta_t(k)}{L}, \end{aligned}$$

where

$\alpha_t(k)$ is the forward probability;

$\beta_t(k)$ is the backward probability.

In order to prevent numerical overflow or underflow issues when the sequence length T just goes moderately large, some scaling techniques are also applied onto $\alpha_t(k)$ and $\beta_t(k)$, in addition to the likelihood L.

Following Equation 3.11, we generate the sequence probabilities $\Pr_t(k)$ in all the hidden states at each time point. This gives a K by T matrix, with row sums satisfying $\sum_{k=1}^{K} \Pr_t(k) = 1$. The probability in the Bull market is plotted in Figure 18. The x-axis represents time, and the y-axis shows the probability in Bull state.

.

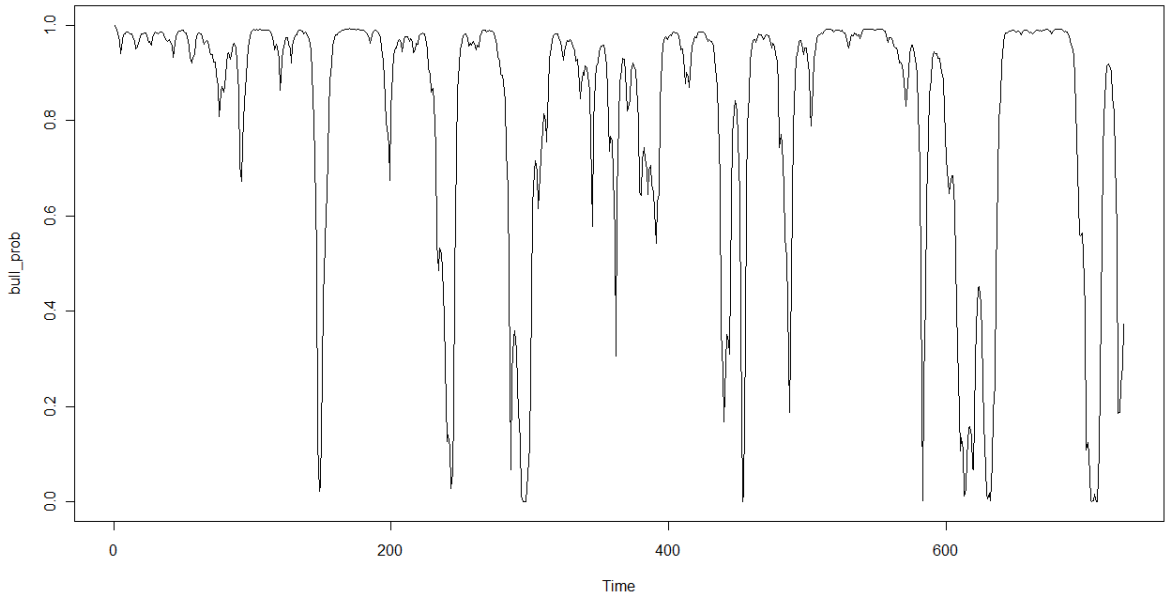

Figure 18. The probability in the Bull market at each time point

3.4.9 Most Likely State Sequence

If we want to explicitly identify the hidden state status at each point, in one way we can assign the data point into the state with maximum local state probability at each time point, in the other way we can get it from the most likely state sequence.

The straightforward way is to calculate each possible state sequence probability given the estimated HMM, and then select the one with the maximum probability. Since the data point can be in K hidden states each time, and the sequence length is T, thus there are K^Tstate sequences probabilities to calculate. It also takes some time to calculate one state sequence probability. Therefore, the naive way to get the most likely state sequence is computationally intensive or even infeasible, depending on K and T.

The Viterbi algorithm (48) (17) (15) was developed to calculate the most likely state sequence in a much quicker way. At each time point, the algorithm only calculates the state probability in each hidden states from the most likely previous sequence path, after doing it for all the time points, it traces back to each time hidden state based on the maximum state probability.

After implementing the Viterbi algorithm, we get the most likely state sequence $MLS(t)$, and plotted in Figure 19:

It is of interest to compare the state sequence between the most likely one and the state probability. The cross-tabulation between two sequences are showed in Table XVIII.

The states sequences from two methods agree with each other in most cases, while the most likely state sequence has more state 2s.

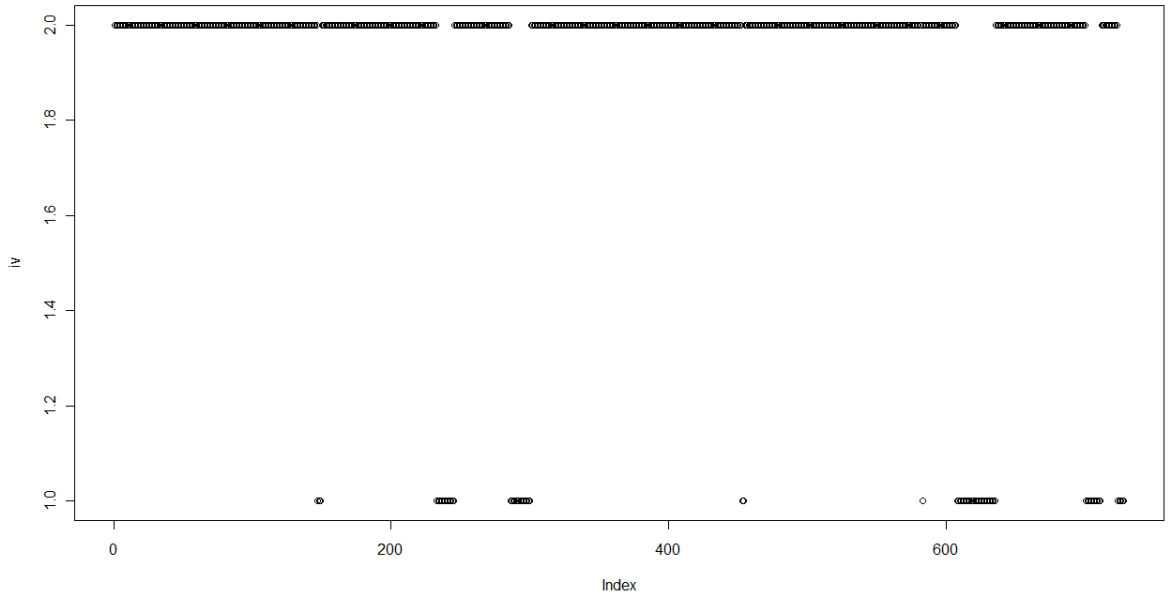

Figure 19. The most likely state sequence of monthly *ROR*

TABLE XVIII

HARD CLASSIFICATION VS. MOST LIKELY SEQUENCE

	Hard Classification		
Most Likely Sequence		State=1	State=2
	State=1	75	3
	State=2	24	626

3.4.10 Forecasting the Future State Sequence

Using the two-state HMM, we can also predict h-step ahead state probabilities $Prob_h(k)$, where $h = T+1, ..., T+h$, and $k = 1, ..., K$. The formula is same as that for an ordinary Markov chain: $Prob_h = Prob_T \times P^(h-T)$, where $Prob_T$ is a $1 \times K$ vector, and P is the $K \times K$ transition probability matrix. The ten-step ahead state probabilities is in Table XIX.

Although the current hidden state is more likely in the state 1, after 1 step, the future hidden states migrates to the state 2.

After we have the prediction of the future hidden state probability, we can further predict the future ROR value, which can be derived from Equation 3.11.

$$ROR_h = Prob_h \times \mu, \tag{3.11}$$

where

$Prob_h$ is a $1 \times K$ vector; μ is a $K \times 1$ vector.

TABLE XIX

TEN-STEP AHEAD STATE PROBABILITY FORECASTING

$h-T$	State 1 Probability	State 2 Probability.
0	0.6265769	0.3734231
1	0.5478667	0.4521333
2	0.4832978	0.5167022
3	0.4303294	0.5696706
4	0.3868776	0.6131224
5	0.3512324	0.6487676
6	0.3219913	0.6780087
7	0.2980037	0.7019963
8	0.2783259	0.7216741
9	0.2621834	0.7378166
10	0.2489411	0.7510589

The result is in Table XX.

Based on the prediction, the monthly ROR of $S\&P500$ will have positive values from January 2011 to July 2011. We hope this provides some sign of current economic recovery around the corner .

3.5 A Simple Coding Method for the Bull and Bear Market

In his blog, Brett Steenbarger, financially-interested psychologist, wrote an article 'Bull and Bear Days: A Simple Coding System and What It Tells Us', where he created a simple definition of Bullish and Bearish days in the market.

Steenbarger defined a day as the Bull one if that day's high is greater than the high of the previous day ($H_t > H_{t-1}$),its low is above the low of the previous day ($L_t > L_{t-1}$), and its close

TABLE XX

TEN-STEP AHEAD MONTHLY *ROR* FORECASTING

h-T	*ROR*
1	-0.00314
2	-0.00156
3	-0.00026
4	0.0008
5	0.001671
6	0.002386
7	0.002972
8	0.003453
9	0.003847
10	0.004171

is bigger than its open ($C_t > O_t = C_{t-1}$). Conversely Steenbarger thought the Bear day should meet three criterions: its high must be below the high of the previous day ($H_t < H_{t-1}$), its low must be below of the low of the prior day ($L_t < L_{t-1}$), and it must show a negative change from open to close ($C_t < O_t = C_{t-1}$). If we give one point for each of the Bullish criteria, then the score of 3 means that day is a Bull day, the score of 2 shows the day is near to the Bull one, the score of 1 represents the day is close to the Bear one, and the score 0 is a clear Bear day. Steenbarger's simple coding for the Bull market can be descried in Equation 3.12.

$$SB_t = I(H_t > H_{t-1}) + I(L_t > L_{t-1}) + I(C_t > C_{t-1}) \tag{3.12}$$

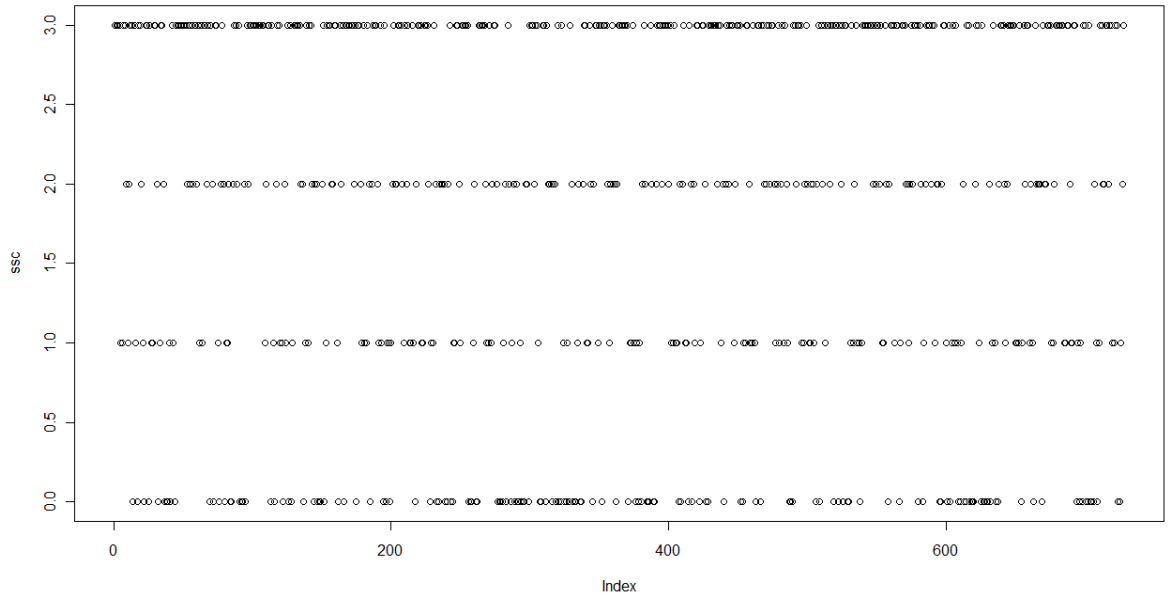

Figure 20. The soft simple coding sequence of monthly ROR

Where

$I()$ is an indication function. Steenbarger's coding system is based on the price change alone, and can be used for any trading instrument as long as there are open, high, low and close price.

Following Steenbarger's definition, and changing the time unit t from day to month, we can code our S&P 500 data into the Bullish score sequence, and call it soft simple coding sequence ($SSC(t)$), because it defines two statuses between the Bull market and the Bear market. The time series plot of $SSC(t)$ is Figure 20.

The frequency table of $SSC(t)$ possible values is Table XXI.

TABLE XXI

THE FREQUENCY TABLE OF $SSC(T)$

Possible Value	Number of Months	Percentage
0 (Bear)	152	21%
1 (Near Bear)	129	18%
2 (Near Bull)	145	20%
3 (Bull)	302	41%
Total	728	100%

Once we consider any score greater than or equal to 2 as the Bull market, then we can have an explicit classification of the market, and we call it hard simple coding sequence ($HSC(t)$). The plot of $HSC(t)$ is in Figure 21.

The frequency table of $HSC(t)$ possible values is in Table XXII.

TABLE XXII

THE FREQUENCY TABLE OF $HSC(T)$

Possible Value	Number of Months	Percentage
0 (Bear)	281	39%
1 (Bull)	447	61%
Total	728	100%

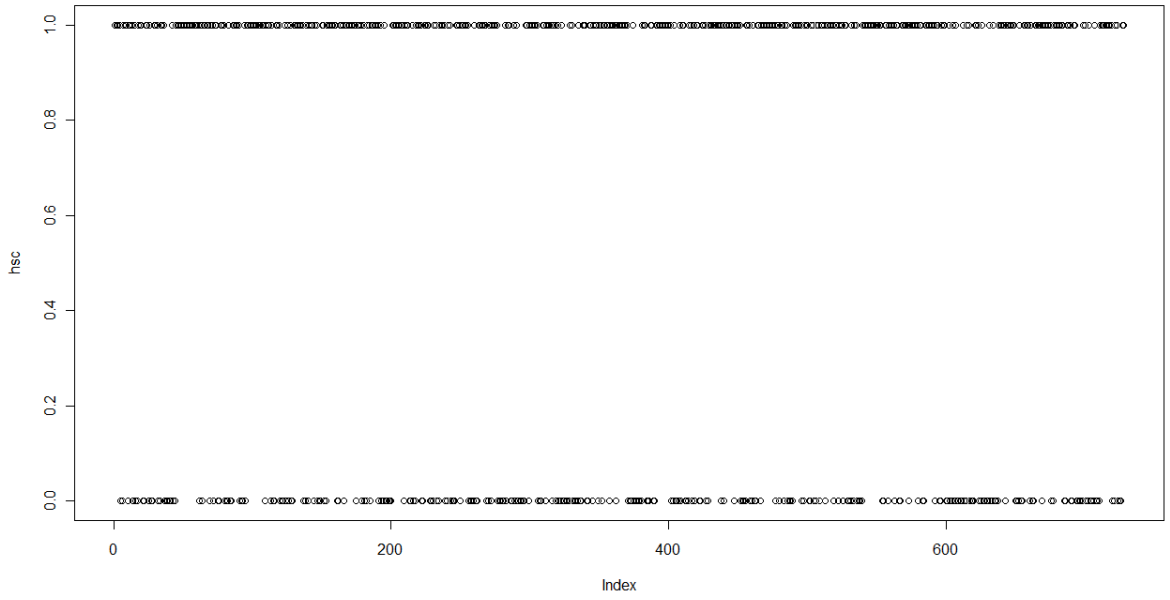

Figure 21. The hard simple coding sequence of monthly ROR

Steenbarger's method of defining the Bull and Bear market is useful and valuable to a certain extent. It incorporates the Bull market price-up trend and the Bear market price-down trend in a straightforward way. It does not require sophisticated mathematical and statistics knowledge, and any spreadsheet software can classify the historical market into the Bull one and the Bear one quickly. While we think it has some limitations. It ignores the possible differences between the Bull market return variance and the Bear market return variance. In financial area, volatility is also a big factor; in some circumstances it is more important than the expected value. It is difficult to use Steenbarger's simple coding method to forecast the near future market segment directly and easily, because the future prices are unknown. One possible way to do that is to estimate future open price, high price, low price, and close price, which could be more changeling than predicting the market segment. In most situations, if an analyst can predict future prices accurately, there will be no need to forecast the future market segment, he can use those forecasts directly to implement some useful trading strategies.

3.6 Comparison of the Simple Coding and the Hidden States

The Bear and Bull market concept defined from the two-state HMM is different from the simple coding method in terms of the methodology. Here we want to check whether the results are also different.

First we will compare the $SSC(t)$ with $Prob_t(k)$, since both of them belong to soft classification. We can look at the average Bull market probability from $Prob_t(k)$ at each possible $SSC(t)$. The result is in Table XXIII.

TABLE XXIII

THE AVERAGE STATE 2 PROBABILITY WITH EACH $SSC(T)$ VALUE

Value of $SSC(t)$	0	1	2	3
Average State 2 Probability	0.652116	0.827264	0.818317	0.891866

Recall that the value '3' in $SSC(t)$ means the Bull market, and '0' represents the Bear market. The Bear market from simple coding has the lowest average state 2 probability, and the Bull market has the highest average state 2 probability. But the average state 2 probability of the near Bear market is just a little bit higher than the one in the near Bull market. The average state 2 probability of the simple coding Bear market is even greater than 0.5. The box plot Figure 22, which comes from the state 2 probability grouped by the simple coding market values, also shows there are no significant state 2 probability differences among the simple coding market values. In other words, there are no relationship between the simple coding market values and the two-state HMM state probability.

The second comparison between the simple coding and the 2-state HMM results can be done by focusing the $HSC(t)$ and the most likely state sequence $MLS(t)$, because both of them are hard classification. The cross-tabulation between $HSC(t)$ and $MLS(t)$ is Table XXIV.

The matching rate between $HSC(t)$ and $MLS(t)$ is only 65.25%, which brings us to think two results are different.

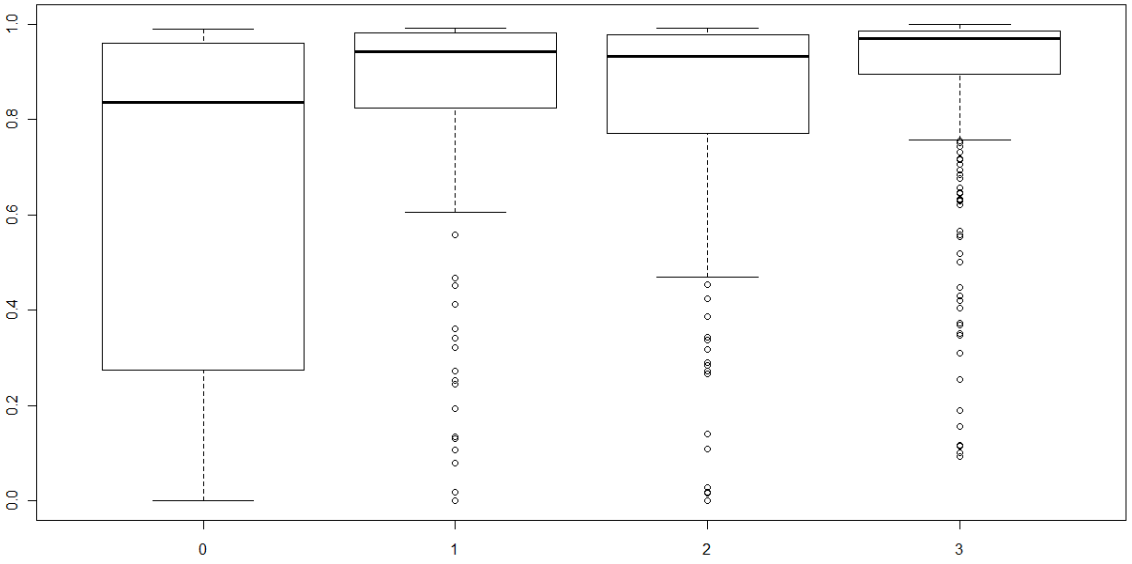

Figure 22. The box-plot of the state 2 probability among the $SSC(t)$ values

TABLE XXIV

CROSS-TABULATION BETWEEN $HSC(T)$ AND $MLS(T)$

	$HSC(t)$	
$MLS(t)$	Bear Market	Bull Market
State 1	53	25
State 2	228	422

In summary, from both soft classification and hard classification results, we conclude that the Bear market and the Bull market definitions from the simple coding method and our two-state HMM are different.

3.7 Summary and Future Research

In this chapter, we focused on applying the HMM to the monthly ROR of S&P 500 data from January 1950 to August 2010. The state-dependent distribution we used is the normal distribution. We tried three HMMs with the number of the hidden states from one to three, and selected the two-state HMM as our final model based on the BIC criterion. For the model estimation, we utilized the MLE method and chose the direct numerical method to optimize the objective function. Following previous research, we implemented some scaling technique to prevent the underflow or the overflow issue. We also created some working parameters, and made certain transformations to get those constrained free parameters. Our two-state HMM estimates can be used to define and describe some characteristics of the Bull market and the Bear market. Our result shows the average ROR in the Bull market is positive, and the counterpart

in the Bear market is negative, which is consistent with the intuition and commonly recognized by different definitions of those two markets. More importantly, our two-state HMM tells us that the variance of ROR in the Bull market is smaller than the one in the Bear market, which confirms the conventional wisdom on this point. Further, we discussed a simple coding method for the Bull market and the Bear market. We compared the results between two methods, which are different. The simple coding method can not predict the further market segment easily, while our two-state HMM can achieve it based on the formulas from the Markov Chain theory and the Finite Mixture Model.

There are some problems left for further research. From the operations research area, a good numeric optimizer maybe can provide better parameter estimates than those in this research. In this research, we assumed the Markov Chain has a homogenous transition probability matrix from time to time, but maybe some kind of non-homogenous transition probability affected by some external variables is more realistic. We only provided the point estimates of the parameters here, while the parameters estimates variances are also useful to test whether the parameters are significant, and whether μ_1 is statistically different from μ_2. The parametric bootstrapping is reported in the current literature for calculating the parameter estimate variances. Our definition for the Bull market and the Bear market is consistent with the current common concept about two markets in terms of the price up-trend and down-trend, in addition we propose that the variance in the two markets are also different. The incorporation of our two markets definition in some trading strategy development, or in investment portfolio optimization, are also interesting research topics.

CHAPTER 4

A SEGMENTATION FOR A CHARITY DONATION DATASET BASED ON THE HIDDEN MARKOV MODEL

4.1 Research Background

With the development of Information Technology (IT), more and more companies established their own customer relationship management (CRM) database. On one hand, more detailed information about customers' behavior is recorded at the transaction level. For example a retail store like Costco may have each customer's transaction information, a credit card company records each cardholder's spending information, and an insurance company has each insured's claim history. On the other hand, it is an important strategy for companies to sell new services or products to the existing customers, because sometimes it is more profitable, compared to selling products to new customers, given tough competition and expensive acquiring costs. Some companies also build some databases to their solicitation activities.

In marketing practice, companies segment their existing customers into different marketing buckets, in order to provide customized products or services to meet their wants or needs. For example, a credit card company may classify some of their customers as Transactors who pay their balance in full every month, and other customers as Revolvers who only pay the amount between the minimum payment and the full payment. Obviously, the two types of the customers have different needs from a credit card. Transactors use credit cards as one of their

payment vehicles, while Revolvers treat credit cards as one of possible financial tools. If a credit card company want to increase their customers usage to gain market share, it should provide some spending rewards for Transactors, while give lower Annual Percentage Rate (APR) to Revolvers. A promotion campaign using low APR will not affect Transactors behavior.

Companies face two challenges in segmenting the existing customers based on their historical behaviors. One is how to perform the segmentation objectively. Given the credit card company example, it is subjective to classify the customers into two categories based on the payment, but why not three or four segments? For the window size of customers payment behavior, is one month a good choice or six months better? The other challenge is that we can not assume that a customer never change his or her needs, and wants; rather, the customer's behavior can change from time to time because of some hidden reasons which are not shown in the available data. It is difficult to predict the customer's future segment, because future behavior for segmentation needs to be forecasted first.

We think that HMM provides a possible solution for this kind of problems, because it segments the data, models the final decision, and allows each case's segment to change following a Markov process. The most important feature of HMM is that it performs the above three tasks simultaneously.

In Netzer's paper, he proposes to use an HMM to model the dynamics of customer relationships through typical transaction data. He also thinks that the HMM can dynamically segment the customer base and examines methods by which the firm can change customers' behavior.

The data he studied is a longitudinal gift-giving dataset, in the context of alumni relations. Finally he demonstrates improved prediction ability on a hold-out sample.

Originally, we wanted to study some transaction and solicitation data from the credit card industry. Unfortunately, due to the sensitivity of each customer's information, we could not find any public datasets. In the general discussion section of Netzer's paper, he mentioned that it would be constructive to investigate the application of HMM in relationship marketing contexts other than the university-alumni relationship. One of the possible applications is other institutional gift giving. We found a newly available public dataset about appeal-donation activities from a US charity; thus we think it is worthwhile to study this dataset with HMM.

4.2 Dataset Introduction

4.2.1 Data Files Exploration

The dataset is obtained from the Direct Marketing Educational Foundation (DMEF), and records information about a leading US charity. It is called dataset 7 in the DMEF, and contains the solicitation and donation history for over one million individual donors. The solicitation history spans 15 years (1992-2006), and the donation history covers 14 years (1993-2006). There are nine data files, four of them in SAS format data, while the remaining are TXT files. After exploring and comparing those two different format data files, we found they provide the same information. Since SAS format data are ready for analysis in the SAS system, our research is based on those four data SAS data files.

The first data file is called '*donor*'. There are 1,176,021 records and two fields in the data, namely donor's ID, and the date when a donor made his first donation. At first we considered

each record representing a distinct donor, but after checking the range of the first donation date, we saw that a very small portion of donors' dates are '1/1/2200', which is far beyond current dates, thus we suspect not all of the people in this dataset ever made a donation, that some of them just had been solicited, but never made any donations.

The second data file is called '*source*'. There are 50,344 observations and two variables in the data, source id and its cost. The data can be used to extract solicitation cost for each donor.

The third file is called '*appeal*', which includes the solicitation history for those potential donors between 1992 and 2006. There are 28,710,911 solicitation records. Three variables in the data are donor's ID, appeal date (solicitation date), and source ID. There are 1,130,037 unique donors targeted according to the data; roughly each donor has been solicited 25 times up to 2006. On the basis of that number, we know that the company does repeatedly solicit its previous donors.

The fourth file is called '*trans*', which contains the donation history across 14 years. There are six variables in the data, including donor's id, donation amount, donation date (gift date), source id, center id and zip code in each observation. Although there are 3,071,050 donation records, the data have only 1,114,478 unique donors, which confirms for us the assumption that some donors do donate more than once.

4.2.2 Pre-Analysis of the Data Files

The analysis we performed is to compare the number of donors among data files '*appeal*', '*trans*' and '*donor*'.

There are more unique donors in the data file '*appeal*' than the data file '*trans*' (1,130,037 vs. 1,114,478), even though the data file '*appeal*' starts later than the data file '*trans*'. The earliest recorded solicitation date is '1/1/1992', while the earliest recorded donation date is '10/31/1991'. In the business sense, we think this difference can be explained by the fact that not all the solicited potential donors ever make a donation.

There are fewer distinct donors in the data file '*appeal*' than the ones of the data file '*donor*' (1,130,037 vs. 1,176,021). There are two possible reasons for this difference. The first is data files' starting time differences, since data file '*appeal*' records solicitation information since '1/1/1992', while data '*donor*' begins to work on '10/31/1991'. The second explanation is that some donors can still make donation even though they were never solicited individually, maybe they were targeted through some kinds of broad media such as TV advertisement or public donation raising campaign, or it was recommended by their friends and relatives.

There are more unique donors in the data file '*donor*' than those in the data file '*trans*' (1,176,021 vs. 1,114,478), which is due to two things. Not all the donors in the data file '*donor*' ever made any donations, thus they are not all recorded in the data file '*trans*'; the earliest donation date in the data file '*trans*' for each donor is not necessarily his or her first donation date.

4.2.3 Some Decisions on the Usage of the Dataset

The above comparisons provide some hints on how to preprocess the data for our research. Our research objective is to apply a HMM for the marketing solicitation activity and response on the existing customers; therefore we make some decisions on how to utilize the dataset.

The target variable: we will predict whether or not an existing donor makes a donation in a month.

The definition of an existing customer: Once a person makes his first donation, we will consider that the person becomes a customer and assign that month as his vintage month. Any solicitation and donation activities after the vintage month will be taken as the marketing treatment and response for the existing donor.

The donation dataset cleaning: In order to assume that the data file '*trans*' records all the donors' first donation behavior to 2006, we will drop those donors whose first donation date in the 'donor' data is not equal to the earliest donation date in the data file '*trans*'.

More than one appeal or donation within a month: If a donor has been solicited or made his donation more than one time in a same month, we will flag that he has been targeted or made donation.

Month on book vs. calendar month: We define month on book as the number of months since a person become an existing donor. If we look at some customers' behaviors monthly by their months on book, a same month on book will be different calendar months, for those customers make their first donations in different calendar months. We can let each customer's vintage month as one of model's predictors, to capture this variation. While, to start with the simple case, and given the sufficient number of each month booked donor, we propose to focus on a group of donors booked in the same month.

4.3 Study Population and Model Objective

After we got the number of new donors by calendar month, we saw there were 39,333 customers on booked in November 2004. Only fourteen of them have not been solicited between December 2004 and November 2006. In addition, approximately 41% of them made at least one donation since December 2004. We exclude 23,011 customers who never make any donations after November 2004 from our study, because those accounts will not provide any information on why a customer make a donation. To control the computing time within a reasonable limit, we random sample 2,000 accounts from the remaining 16,322 customers, which represents a 12% random sample. There are 24 performance months from December 2004 to November 2006, thus we will use the first 21 month data of each customers to build models, and reserve the last 3 month data as a hold-out part for model prediction evaluation. Although we only choose 2,000 customers, there are actually 42,000 data points for fitting the model, because each customer has 21 months of data points.

The computing facility we used is Dell PowerEdge Server 1900, with Xeon 5000 CPU and 8 GB RAM. For one set of initial values, it takes two hours, twenty-four hours, and one hundred and fourteen hours to obtain a resolution from two-state, three-state and four-state HMMs respectively.

We want to study the relationship between appeal and donation through models, and therefore whether or not making a donation in a performance month is the target variable, and whether or not being solicited in the same month is the predictor.

```
Call:
glm(formula = Y_train ~ X1_train, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5049  -0.5049  -0.5049  -0.2299   2.7005

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.62004    0.05432  -66.64   <2e-16 ***
X1_train     1.62465    0.05729   28.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25404  on 41999  degrees of freedom
Residual deviance: 24220  on 41998  degrees of freedom
AIC: 24224

Number of Fisher Scoring iterations: 6
```

Figure 23. The Logit model estimation

4.4 Logit Model for the Dataset

We use the Logit model as a benchmark in this study, since it is widely used to model binary outcomes. Using the 'glm' package in R, with making a donation as the dependent variable and receiving an appeal as the independent variable, we obtain the estimation results in Figure 23. The results show an appeal does increase the donation probability because of the positive sign of the predictor. After we transform the estimate to the success rate, it shows the donation probability is 0.02608306 without an appeal, and with an appeal it increases to 0.1196878.

4.5 HMMs for the Dataset

4.5.1 Construction of the HMM

Within the framework of HMM, we think the data can be modeled in the following way:

(1) There are different donation possibility segments, and the donation probabilities are different among the segments;

(2) A customer can stay in the same segment or change to any other segment from time to time, because of different unknown reasons;

(3) A customer donation probability is decided by its underlying segment status;

(4) The appeal activity affects a customer's segment changes, in other words, it has some impact on a customer's transition probability matrix.

Since there are $N = 2000$ customer monthly data or time series, the likelihood function becomes Equation 4.1.

$$P(X|\pi, P, F) = \prod_{i=1}^{N}(\pi F_{1i} \sum_{t=2}^{T} P_{ti} F_{ti}), \tag{4.1}$$

where

π is a 1 by K initial probability vector of all the customers

F_{ti} is a K by K state-dependent density matrix of customer i at time t

P_{ti} is a K by K transition probability matrix of customer i at time t

For the sake of model parsimony, we assume all the customers start from the same initial state instead of different ones, otherwise, there are too many free parameters to be estimated.

Following the same methodology in the previous chapter, we create some working parameters during the optimization procedure, to limit the free parameters with their ranges. The details are described below:

(1) The elements of the initial probability vector π are modeled as:

$$\pi_i = \frac{\exp(\alpha_{\pi_i})}{1 + \sum_{j=1}^{k-1} \exp(\alpha_{\pi_j})}$$

for $i = 1, \ldots, k-1$, and

$$\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$$

(2) The success rate f_i is modeled as:

$$f_1 = \frac{\exp(\alpha_{f_1})}{1 + \exp(\alpha_{f_1})}$$

and

$$f_i = \frac{\exp(\alpha_{f_1} + \sum_{j=1}^{i-1} \exp(\alpha_{f_j}))}{1 + \exp(\alpha_{f_1} + \sum_{j=1}^{i-1} \exp(\alpha_{f_j}))}$$

for $i = 2, \ldots, k$, and Bernoulli probability mass function

$$F_i = f_i^y \times (1 - f_i)^{(1-y)}$$

Here y is the dependent variable, and we pre-define $f_1 \leq f_2 \leq \ldots \leq f_k$, to label the segment.

(3) The elements of the transition probability matrix P are modeled as:

$$p_{ij} = \frac{\exp(\mu_{p_{ij}} + \alpha_{p_{ij}} x)}{1 + \sum_{j=1}^{k} \exp(\mu_{p_{ij}} + \alpha_{p_{ij}} x)}$$

and

$$p_{ii} = 1 - \sum_{j=1}^{k} p_{ij}$$

where $i = 1, \ldots, k$,$j = 1, \ldots, k$ and $i \neq j$.

We allow transitions between any two states, which is different from Netzer's research, where he allows transitions only between adjacent states or to the lowest state.

4.5.2 Three HMM Comparisons

We try the number of hidden states K from 2 to 4, build three HMMs, and estimate them by maximum likelihood estimation. The optimizer we use is nlm() in R.

For the two-state HMM, we use twenty-four different initial values, in order to increase the chance of finding the global optimum. We find the one with 0 for all the working parameters gives a best likelihood. For the three-state and four-state HMMs, we notice it takes more than twenty-four hours, and more than forty-eight hours for a solution with a starting value respectively, therefore we set all of their working parameters initial values as zero.

A scaling technique similar to that in the preceding chapter is adopted, to prevent the numerical underflow here. Finally the log likelihood value and BIC of the three HMMs are listed in Table XXV. With more hidden states, a model gives larger likelihood value, while

more parameters are needed. We choose the three-state HMM as our best model, given that it has the smallest BIC. It is good to know that posterior probability of model $k \propto \exp(-BIC_k/2)$, where $k = 1, 2, \ldots, K$, that is

$$\begin{aligned} \text{posterior probability of model } k &= \frac{\exp(-BIC_k/2)}{\sum_{j=1}^{K} \exp(-BIC_j/2)} \\ &= \frac{\exp[(-BIC_k - BIC_1)/2]}{\sum_{j=1}^{K} \exp[(-BIC_j - BIC_1)/2]} \end{aligned}$$

Given that we have to select one model from three HMMs, we can calculate the posterior probability of choosing a model by using the above formula.

TABLE XXV

BIC OF THREE HMMS

Number of Components	Log Likelihood	Number of Parameters	BIC	Model Prob.
2	-11965.18	7	24004.88	5.574852e-09
3	-11892.91	17	23966.79	1
4	-11886.89	31	24103.79	1.781673e-30

4.5.3 Three-state HMM Estimates

After transforming the working parameters back to the original HMM parameter, we can get the three-state HMM estimate as below:

$$\hat{\pi} = (0, 0.7478418, 0.2521582)$$

$$\hat{P}_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0.1702819 & 0.8297181 & 0 \\ 0.7242869 & 0 & 0.2757131 \end{pmatrix}$$

$$\hat{P}_1 = \begin{pmatrix} 0 & 0.6542168 & 0.3457656 \\ 0 & 0.0000003 & 0.9999917 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\hat{f}_1 = 0.000002315987$$

$$\hat{f}_2 = 0.001266105$$

$\hat{f}_3 = 0.1445806$ where $\hat{P}_0$ is the transition probability without a solicitation, and $\hat{P}_1$ is the one with an appeal.

With regard to the success rate f in different segments, we can describe the state 1 as the impossible donation segment, the state 2 as the possible donation segment, and the state 3 as the likely donation segment. We still define the state 2 as the possible donation segment, based on the experience from the credit card industry, 0.12% is a reasonable result from its regular campaign without significant promotions. Also because of the imbalance between the solicitation cost and revenue, 0.12% should not be treated as a impossible probability. If we

think the mail cost is \$0.30, once the average donation is more than \$250, it is a break-even for 0.12% donation rate; for 14.5% donation rate, more than \$20 per donation covers the cost.

The donors starts from the state 2 with the probability 0.75 , and the state 3 with the probability 0.25. We think the result sensible in that all those donor just make their fist donations in the previous month, it is difficult to explain whey they immediately become impossible to make any future donations without any additional information.

Comparing $\hat{P}_0$ with $\hat{P}_1$, we think the appeal increases the donation probability in the following way:

(1) If the customer's previous state is 1, a solicitation enables it jump to the state 3, instead of just jump to the state 2.

(2) If the customer's previous state is 2, an appeal will increase its chance to the state 3, by reducing its chance to the state 1 or unchange.

(3) If the customer's previous state is 3, a solicitation is important for it to stay in the preferable status, which suggests that a company should keep soliciting its active customers, instead of contacting them only occasionally to avoid bothering them.

4.6 The Link between the HMM and the Logit Model

It is of interest to consider whether there are some links between the Logit model and the HMM. Recall that from the Logit model result, we know the donation probability is 0.02608306 without an appeal, and 0.1196878 with an appeal. Therefore we can consider the Logit model as a two-state HMM, and the parameter estimates are as below:

$$\hat{P}_0 = \begin{pmatrix} 1 & 0 \\ \\ 1 & 0 \end{pmatrix}$$

$$\hat{P}_1 = \begin{pmatrix} 0 & 1 \\ \\ 0 & 1 \end{pmatrix}$$

$\hat{f}_1 = 0.02608306$

$\hat{f}_2 = 0.1196878$

π_1= the percentage of accounts who are not solicited in the first month

π_2= the percentage of accounts who are solicited in the first month

It is hard to compare the log likelihood values between the Logit model and the HMM directly, to evaluate the model fit, since the HMM one is a scaled log likelihood value. While because our two-state HMM estimates are different from the above one, we expect the Logit model fit to be worse than our two-state HMM, and also is worse than the three-state HMM.

To test whether the above inference is valid, we utilize the parameter estimates from the two-state HMM which is equivalent to the Logit model, and calculate the log likelihood value with the same scaling technique. The scaled log likelihood value is -12109.91, which is smaller than those of the HMMs we have in the previous section.

4.7 Out-of-Sample Predictions of the HMM and the Logit Model

We use the reserved three month data for two models' prediction performance comparison. For the Logit model, we apply the model estimation on the reserved sample to get each donor's donation probability. For the three-state HMM, first we use the same formula as in the previous chapter to get each customer's probabilities in different states at the last month in the training

data, then predict its probabilities in different states for those new three months, and finally calculate its donation probability from the weighted sum of each state donation probability, here the state probability is the weight.

The donation activity is a rare event in the test data, for example, if we consider those three month data as a whole part, the donation rate is 0.07216667, and once we look at the donation rate within each month, the first month is 0.095, the second month is 0.096, and the third month is 0.03. So the model results will be useless, if we try to predict donation or no donation, because all the predictions will be 0 by choosing 0.5 as a cut-off point. If we want to choose a lower cut-off, an appropriate one would be industry-specific.

In practice, the marketer considers the predicted occurrence probability as a model score, and uses it to rank the potential customers. Here we adopt this methodology, rank the test sample by both the Logit model prediction and the three-state HMM forecasting, then calculate the cumulative donation rates from those two models. We first rank the three month data together, and the cumulative donation rates are plotted in Figure 24. We further try to rank the donors in the same month separately, and the first month cumulative donation rates are plotted in Figure 25, those of the second month are in Figure 26, and the counterparts of the third month are in Figure 27.

From the results, we notice that when ranking the three months data together, the HMM performs better only in the first 900 cases out of 6,000 data points, which represent 2,000 unique donors. While ranking those 2,000 customers within the same month, the HMM yields better results in the first 1,000 cases out of 2,000 data points very month. We think this is

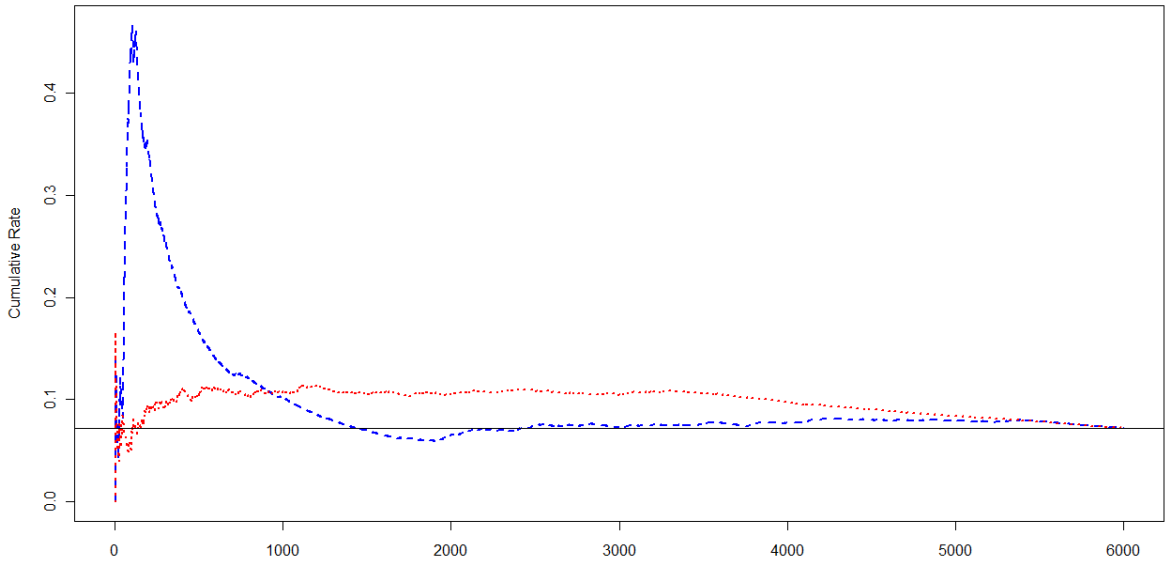

Figure 24. The cumulative donation rate of the HMM and the Logit model with all three months together: the blue line is HMM result, and the red line is Logit model result.

due to the fact that the actual donation rate in the third month suddenly drops, to remove the month trend impact on the model prediction performance, ranking the customers in the same month is more sensible and practicable. Therefore, we think the three-state HMM ranks the customers better than the Logit model does, and we conclude that the three-state HMM has better prediction performance.

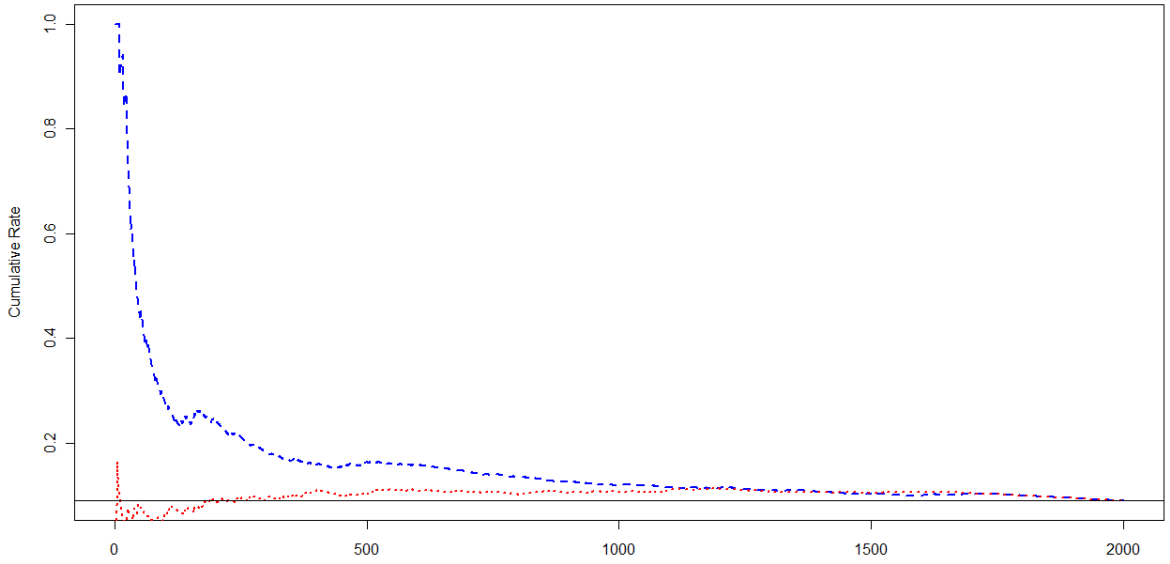

Figure 25. The cumulative donation rate of the HMM and the Logit model for the first forecasting month: the blue line is HMM result, and the red line is Logit model result.

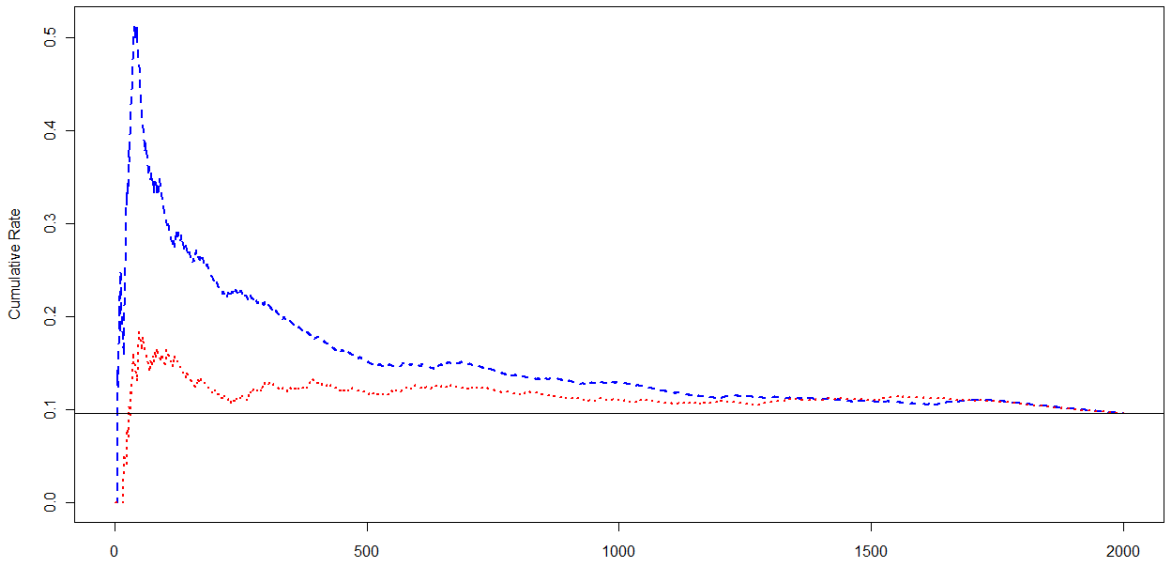

Figure 26. The cumulative donation rate of the HMM and the Logit model for the second forecasting month: the blue line is HMM result, and the red line is Logit model result.

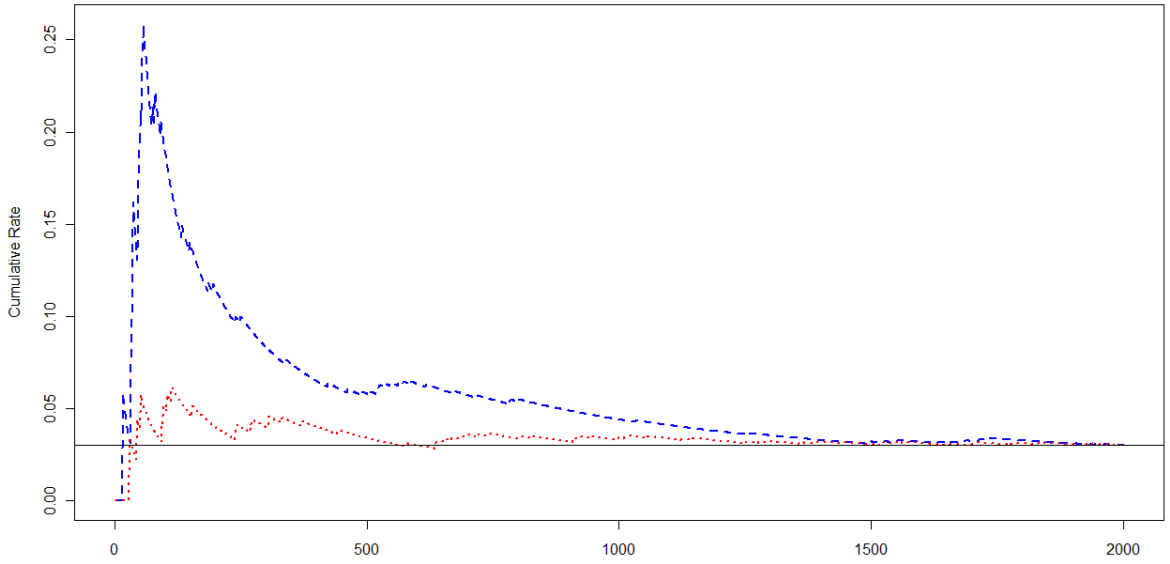

Figure 27. The cumulative donation rate of the HMM and the Logit model for the third forecasting month: the blue line is HMM result, and the red line is Logit model result.

4.8 Summary and Future Research

In this chapter, we studied a charity appeal and donation data from the Direct Marketing Education Foundation. We tried to predict whether or not an existing donor makes a donation as a function of whether or not he is solicited. A Logit model has been built as a benchmark, and it shows that an appeal will increase the donation probability from 0.026 to 0.120. We adopt the HMM framework, and propose that there are different hidden segments for a donor to make donation within which the donation probabilities are different. The appeal activity has an impact on the transition probability matrix based on which a donor can remain or change his segment from time to time. The number of hidden states from two to four is tried, and a three-state HMM is chosen on the basis of its smallest BIC value. We also transfer the logit model to a two-state HMM, and infer its data fit is not as good as our HMM. We prove it by calculating the Logit model scaled log likelihood value by treating it as a two-state HMM. Finally, we use the HMM and the Logit model to score and rank the reserved test data, and show the HMM provides a better out sample prediction from the cumulative donation plots.

This research is our starting point for studying the longitudinal data using the HMM; therefore many things can be done in future research. First, we can further model the initial probability using some customer on book information, for this dataset, the first donation amount can be a good candidate. In addition to model the transition probability matrix, we can also use some predictors to model the state-dependent success rate f simultaneously. We may find a way to study the donors from different vintage month together, instead of just focusing on a particular group. Once the HMM becomes complicated, we can also use a more sophisticated

Logit model as the benchmark, in which the variables about recency, frequency and monetary amount are used as the predictors. Another important research enhancement is to find a more powerful computing facility and method, in order to estimate a complicated HMM within an acceptable timeline.

CHAPTER 5

OVERALL SUMMARY

In this dissertation, we have applied what we have termed *labeling models,* that is, models involving pairs (Y, S) where Y is observable and may be a vector, and S is an unobservable state. The values of S may be identified by $1, 2, \ldots, K$, and are called *labels.* The value K, the number of states, may be estimated by model-selection criteria such as the Bayesian Information Criterion (BIC).

For cross-sectional data, we write (Y_i, S_i), $i = 1, 2, \ldots, n$; for time-series, (Y_t, S_t), $t = 1, 2, \ldots, n$.

In cross-sectional data, the states are clusters of segments of individuals (consumers) and the procedure of estimating the states is called *cluster analysis*, *clustering*, or *market segmentation.* An underlying model is the Finite Mixture Model (FMM), the elements of which are class-conditional p.d.f.s and their prior probabilities ("mixing" probabilities).

An important extension of the FMM is *clusterwise regression.* Here, the observable vector Y is partitioned into response (dependent) variables and explanatory (independent) variables. The relationships between and among them may differ across segments. The regression is logistic regression if the response variable is a binary scalar. The clusterwise logistic regression model is called the Mixture Logit Model (MLM) or Mixture Logistic Regression (MLR). This model was applied to a DMEF dataset to determine characteristics associated with ordering from the firm's website versus ordering from the firm's mailed catalog.

For time series, we write (Y_t, S_t), $t = 1, 2, \ldots, n$. The states S_t correspond to regimes or phases such as Bull and Bear in the market (or, in other work, Recession, Recovery, Expansion, Contraction in macroeconomics.) A segment is a sequence of successive values of t for which the state remains constant. In the thesis we segmented a series of S&P500 monthly rates of return. We found that $K = 2$ states was better than $K = 3$ or $K = 1$. The means of state-conditional Normal distributions were positive for one state and negative for the other, corresponding to conventional notions of Bull and Bear markets.

The hidden Markov model (HMM) is a dynamic FMM. The FMM has p.d.f.

$$f(y) \;=\; \lambda_1 \, f_1(y) \;+\; \lambda_2 f_2(y) + \ldots + \lambda_k f_k(y) + \ldots + \lambda_K f_K(y).$$

Analogously, the HMM has conditional p.d.f. of the form

$$f(y_t | S_{t-1} = j) \;=\; p_{j1} f_1(y_t) + p_{j2} f_2(y_t) + \ldots + p_{jk} f_k(y_t) + \ldots + p_{jK} f_K(y_t),$$

with transition probabilities p_{jk} in the HMM replacing mixing probabilities λ_k in the FMM.

The HMM was applied to a DMEF dataset on charitable donations. The transition probabilities were modeled as a logistic function of a covariate relating to donor solicitation. Future research could involve the use of more covariates and would be facilitated by faster hardware with more memory. On the theoretical side, given an underlying process for the covariates, the convergence of the overall HMM, with covariates integrated out, could be investigated, as could the nature of the convergence of the sequence of parameter estimates.

APPENDICES

Appendix A

AN EM ALGORITHM FOR ESTIMATING THE MIXTURE LOGISTIC REGRESSION

A.1 The Likelihood of the Mixture Logistic Regression

The probability mass function (p.m.f.) of the k-th component is

$$f_k(y) = f(y, \beta_k).$$

Index the components by $k = 1, 2, \ldots, K$ and the observations by $i = 1, 2, \ldots, n$. Let

$$p_{ki} = \frac{\exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)} = \frac{1}{1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)}.$$

and

$$q_{ki} = 1 - p_{ki} = \frac{1}{1 + \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)} = \frac{\exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)}{1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)}.$$

Let

$$f_{ki} = p_{ki}^{y_i} \, q_{ki}^{1-y_i}.$$

The p.m.f. of Y_i is

$$f_i = \sum_{k=1}^{K} \lambda_k f_{ki}.$$

Appendix A (Continued)

The likelihood is

$$L = \prod_{i=1}^{n} f_i.$$

A.2 Maximizing the Likelihood

Usually one takes the log likelihood, but in this case let's examine the likelihood itself.

Remember that

$$\sum_{k=1}^{K} \lambda_k = 1.$$

Introduce a Lagrange multiplier λ; then the maximization problem is to maximize

$$M = L + \lambda(1 - \sum_{k=1}^{K} \lambda_k) = \prod_{i=1}^{n} f_i + \lambda(1 - \sum_{k=1}^{K} \lambda_k).$$

$$\frac{\partial M}{\partial \lambda_k} = \frac{\partial L}{\partial \lambda_k} + \lambda.$$

$$\frac{\partial fi}{\partial \lambda_k} = \frac{\partial}{\partial \lambda_k} \sum_{j=1}^{K} \lambda_j f_{ji} = \sum_{j=1}^{K} \frac{\partial}{\partial \lambda_k} \lambda_j f_{ji} = f_{ki}.$$

Appendix A (Continued)

The likelihood is $L = \prod_{i=1}^{n} f_i$. Let θ denote a parameter, either λ or β_{kv}. Then

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \frac{\partial \prod_{i=1}^{n} f_i}{\partial \theta} \\
&= \sum_{i=1}^{n} \frac{\partial f_i}{\partial \theta} \prod_{j \neq i} f_j \\
&= \sum_{i=1}^{n} [(\partial f_i / \partial \theta)/f_i) \prod_{j=1}^{n} f_j \\
&= \sum_{i=1}^{n} [(\partial f_i / \partial \theta)/f_i] \, L \\
&= L \sum_{i=1}^{n} [(\partial f_i / \partial \theta)/f_i].
\end{aligned}$$

A.2.1 Derivative with respect to λ_k

Then

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_k} &= L \sum_{i=1}^{n} [(\partial f_i / \partial \lambda_k)/f_i] \\
&= L \sum_{i=1}^{n} (f_{ki}/f_i) \\
&= L \sum_{i=1}^{n} (1/\lambda_k)(\lambda_k f_{ki}/f_i) \\
&= L \sum_{i=1}^{n} (1/\lambda_k) \Pr(\Pi_k | y_i, \boldsymbol{x}_i) \\
&= (1/\lambda_k) L \sum_{i=1}^{n} \Pr(\Pi_k | y_i, \boldsymbol{x}_i)
\end{aligned}$$

Appendix A (Continued)

We have $\partial M/\partial\lambda_k = \partial L/\partial\lambda_k - \lambda$. Setting this equal to 0 gives

$$\partial L/\partial\lambda_k = \lambda,$$

or

$$(1/\lambda_k)L\sum_{i=1}^{n}\Pr(\Pi_k|y_i,\boldsymbol{x}_i) = \lambda.$$

Solving for λ_k gives

$$(1/\lambda)L\sum_{i=1}^{n}\Pr(\Pi_k|y_i,\boldsymbol{x}_i) = \lambda_k.$$

or

$$(L/\lambda)\sum_{i=1}^{n}p(k|y_i,\boldsymbol{x}_i) = \lambda_k,$$

where

$$p(k|y_i,\boldsymbol{x}_i) = \Pr(\Pi_k|y_i,\boldsymbol{x}_i).$$

Since $\sum_{k=1}^{K}p(k|y_i,\boldsymbol{x}_i) = 1$ and $\sum_{k=1}^{K}\lambda_k = 1$, summing gives $\lambda = Ln$, and the MLE of λ_k is

$$\hat{\lambda_k} = (1/n)\sum_{i=1}^{n}p(k|y_i,\boldsymbol{x}_i).$$

A.2.2 Derivatives with respect to the regression coefficients

Now,

$$\text{logit }(\pi_k) = \boldsymbol{\beta}_k'\boldsymbol{x},$$

Appendix A (Continued)

where the first component of $\boldsymbol{x}$ is 1 and the first component of $\boldsymbol{\beta}_k$ is an intercept $\beta_{0k} = \alpha_k$. Index the components by $k = 1, 2, \ldots, K$, and the explanatory variables by $v = 0, 1, 2, \ldots, p$. The partial of the likelihood with respect to β_{kv} is

$$\begin{aligned}
\frac{\partial L}{\partial \beta_{kv}} &= \partial \prod_{i=1}^{n} f_i / \partial \beta_{kv} \\
&= \sum_{i=1}^{n} \partial f_i / \partial \beta_{kv} \prod_{j \neq i} f_j \\
&= \sum_{i=1}^{n} \partial f_i / \partial \beta_{kv} \, (1/f_i) [\prod_{j=1}^{n} f_j] \\
&= \sum_{i=1}^{n} \partial f_i / \partial \beta_{kv} \, (1/f_i) \, L \\
&= L \sum_{i=1}^{n} (1/f_i) \partial \sum_{j=1}^{K} \lambda_j f_{ji} / \partial \beta_{kv} \\
&= L \sum_{i=1}^{n} (1/f_i) \lambda_k \partial f_{ki} / \partial \beta_{kv}
\end{aligned}$$

where

$$f_i = \sum_{k=1}^{K} f_{ki}, \qquad f_{ki} \; = \; p_{ki}^{y_i} \, (1 - p_{ki})^{1 - y_i},$$

and

$$p_{ki} \; = \; \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i) / [1 + \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)] \; = \; 1 / [1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)].$$

Appendix A (Continued)

A.2.3 Various partial derivatives

We have

$$\frac{\partial f_i}{\partial \beta_{kv}} = \frac{\partial}{\partial \beta_{kv}} \sum_{j=1}^{K} \lambda_j f_{ji} = \lambda_k \frac{\partial f_{ki}}{\partial \beta_k}.$$

Remember that $f_{ki} = p_{ki}^{y_i} (1 - p_{ki})^{1-y_i}$. We have to differentiate p_{ki}. Continuing,

$$\begin{aligned}
\frac{\partial p_{ki}}{\partial \beta_{kv}} &= \frac{\partial}{\partial \boldsymbol{\beta}_{kv}} \frac{\exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}_k' \boldsymbol{x}_i)} \\
&= \frac{\partial}{\partial \beta_{kv}} [1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)]^{-1} \\
&= (-1)[1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)]^{-2} \frac{\partial}{\partial \beta_{kv}} [1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)] \\
&= (-1)[1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)]^{-2} \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)(-x_{ki}) \\
&= [1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)]^{-2} \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i) \\
&= x_{ki} \{1/[1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i]\} \{\exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i)/[1 + \exp(-\boldsymbol{\beta}_k' \boldsymbol{x}_i]\} \\
&= x_{ki}\, p_{ki}\, q_{ki}\,.
\end{aligned}$$

Appendix A (Continued)

Now we use this result to calculate the partial of f_{ki}.

$$\begin{aligned}
\frac{\partial f_{ki}}{\partial \beta_{kv}} &= \frac{\partial}{\partial \beta_{kv}} p_{ki}^{y_i}\, q_{ki}^{1-y_i} = \frac{\partial}{\partial \beta_{kv}} p_{ki}^{y_i}\,(1-p_{ki})^{1-y_i} \\
&= [\frac{\partial}{\partial \beta_{kv}} p_{ki}^{y_i}][q_{ki}^{1-y_i}] + [p_{ki}^{y_i}][\frac{\partial}{\partial \beta_{kv}}\,(1-p_{ki}^{1-y_i}) \\
&= [\frac{\partial}{\partial \beta_{kv}} p_{ki}^{y_i}][q_{ki}^{1-y_i}] - [p_{ki}^{y_i}][\frac{\partial}{\partial \beta_{kv}}\, p_{ki}^{1-y_i}] \\
&= y_i[p_{ki}^{y_i-1}][\frac{\partial}{\partial \beta_{kv}}\, p_{ki}]q_{ki}^{1-y_i} \;+\; p_{ki}^{y_i}(1-y_i)(1-p_{ki})^{-y_i}[\frac{\partial}{\partial \beta_{kv}} p_{ki}] \\
&= y_i\, p_{ki}^{y_i-1} q_{ki}^{1-y_i}(\partial p_{ki}/\partial \beta_{kv}) - p_{ki}^{y_i} q_{ki}^{-y_i}(\partial p_{ki}/\partial \beta_{kv}) + y_i p_{ki}^{y_i} q_{ki}^{-y_i}(\partial p_{ki}/\partial \beta_{kv}) \\
&= p_{ki}^{y_i-1}\, q_{ki}^{-y_i}(\partial p_{ki}/\partial \beta_{kv}) \cdot [y_i q_{ki} - p_{ki} + y_i p_{ki}] \\
&= p_{ki}^{y_i-1} q_{ki}^{-y_i}(\partial p_{ki}/\partial \beta_{kv}) \cdot (y_i - p_{ki}) \\
&= p_{ki}^{y_i-1}\, q_{ki}^{-y_i}(x_{vi} p_{ki} q_{ki})(y_i - p_{ki}) \\
&= x_{vi}\, p_{ki}^{y_i-1}\, q_{ki}^{-y_i}(x_{vi} p_{ki} q_{ki})(y_i - p_{ki}) \\
&= x_{vi}\, p_{ki}^{y_i}\, q_{ki}^{1-y_i}\,(y_i - p_{ki}) \\
&= x_{vi}\, f_{ki}(y_i - p_{ki}).
\end{aligned}$$

Now use this result to calculate the partial of f_i.

$$\frac{\partial fi}{\partial \boldsymbol{\beta}_k} \;=\; \frac{\partial}{\partial \boldsymbol{\beta}_k} \sum_{j=1}^{K} \lambda_j f_{ji} \;=\; \sum_{j=1}^{K} \frac{\partial}{\partial \boldsymbol{\beta}_k} \lambda_j f_{ji} \;=\; \lambda_k \frac{\partial}{\partial \boldsymbol{\beta}_k} f_{ki} \;=\; \lambda_k\, x_{vi}\, f_{ki}(y_i - p_{ki}).$$

Appendix A (Continued)

Now the partial of L can be computed.

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\beta}_{kv}} &= L\sum_{i=1}^{n}\frac{\partial f_i/\partial\beta_{kv}}{f_i}\\ &= L\sum_{i=1}^{n} x_{vi}(\lambda_k f_{ki}/f_i)(y_i - p_{ki})\\ &= L\sum_{i=1}^{n} x_{vi}\, p(k|y_i,\boldsymbol{x}_i)\,(y_i - p_{ki}).\end{aligned}$$

The equation $\partial L/\partial\beta_{kv} = 0$ is equivalent to

$$\sum_{i=1}^{n} x_{vi}\, p(k|y_i,\boldsymbol{x}_i)\,(y_i - p_{ki}) \;=\; 0,$$

or

$$\sum_{i=1}^{n} x_{vi}\, p(k|y_i,\boldsymbol{x}_i)\, y_i \;=\; \sum_{i=1}^{n} x_{vi}\, p(k|y_i,\boldsymbol{x}_i)\, p_{ki}.$$

Note the central role played by the posterior probabilities of group membership, $p(k|y_i,\boldsymbol{x}_i)$. Note that $p(k|y_i,\boldsymbol{x}_i) = \lambda_k p_{ki}^{y_i}\, q_{ki}^{1-y_i}/f_i$.

In particular, this involves f_i, which involves all the components, not just the k-th.

CITED LITERATURE

1. M.R. Anderberg. *Cluster Analysis for Applications (Probability & Mathematical Statistics Monograph)*. Academic Press Inc, New York, 1974.

2. Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

3. Peter Boatwright, Robert McCulloch, and Peter Rossi. Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *Journal of the American Statistical Association*, 94(448):1063–1073, 1999.

4. George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

5. G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of American Statistical Association*, 95: 957–970, 2000.

6. Jiahua Chen and J.D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24(2):167–175, 1996.

7. Rong Chen and Thomas B. Fomby. Forecasting with stable seasonal pattern models with an application to Hawaiian tourism data. *Journal of Business & Economic Statistics*, 17(4):497–504, 1999.

8. Francesca Chiaromonte, R. Dennis Cook, and Bing Li. Sufficient dimension reduction in regressions with categorical predictors. *Annals of Statistics*, 30(2):475–497, 2002.

9. Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

10. N. Chopin. Inference and model choice for sequentially ordered hidden Markov models. *Journal of the Royal Statistical Society*, B(69):269–284, 2007.

11. Hwan Chung, Eric Loken, and Joseph L. Schafer. Difficulties in drawing inferences with finite-mixture models: A simple example with a simple solution. *The American Statistician*, 58(2):152–158, 2004.

12. Gilbert A. Churchill and Dawn Iacobucci. *Marekting Research: Methodological Foundations. 9th ed.* Cengage South-Western, 2004.

13. Meghana Deodhar and Joydeep Ghosh. A framework for simultaneous co-clustering and learning from complex data. Technical report, Department of Electrical and Computer Engineering, The University of Texas at Austin, 2007.

14. Wayne S. DeSarbo and William L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5:249–282, 1988.

15. Rakesh Dugad and U.B. Desai. A tutorial on hidden Markov models. Research memorandum, Department of Electrical Engineering, Indian Institute of Technology, 1996.

16. B.S. Everitt and D.J. Hand. *Finite Mixture Dsitributions.* Chapman and Hall, London, 1981.

17. G.D. Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

18. Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F. M. Smith. Illustration of Bayesian inference in Normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 84(412):972–985, 1990.

19. Bettina Grün and Friedrich Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics and Data Analysis*, 2006.

20. Markus Hahn and Jorn Sass. Parameter estimation in continuous time Markov switching models: A semi-continuous markov chain monte carlo approach. *Bayesian Analysis*, 4(1):63–84, 2009.

21. John Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

22. Shane T. Jensen, Blakeley B. McShane, and Abraham J. Wyner. Hierarchical Bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, 4(4):631–652, 2009.

23. S. John. On identifying the population of origin of each observation in a mixture of observations from two Normal populations. *Technometrics*, 12:553–563, 1969.

24. S. John. On identifying the population of origin of each observation in a mixture of observations from two Gamma populations. *Technometrics*, 12:565–568, 1970.

25. Sougata Kerr and Lucia Dunn. Consumer search behavior in the changing credit card market. *Journal of Business and Economic Statistics*, 26(3):345–353, 2008.

26. Philip Kotler and Gary Armstrong. *Principles of Marketing. 13th ed.* Prentice Hall, Englewood Cliffs, NJ.

27. Philip Kotler and Kevin Lane Keller. *Marketing Management, 13th ed.* Prentice Hall, Englewood Cliffs, NJ.

28. Yee Leung, Jiang-Hong Ma, and Wen-Xiu Zhang. A new method for mining regression classes in large data sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):5–21, 2001.

29. Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

30. Gary L. Lilien, Arvind Rangaswamy, and Arnaud De Bruyn. *Principles of Marketing Engineering*. Trafford Publishing, 2007.

31. Asger Lunde and Asger Timmermann. Duration dependence in stock prices:an analysis of Bull and Bear markets. *Journal of Business & Economic Statistics*, 22(3):253–273, July 2004.

32. Lan Luo, P.K. Kannan, and Brian T. Ratchford. Incorporating subjective characteristics in product design and evaluations. *Journal of Marketing Research*, pages 182–194, 2008.

33. John M. Maheu and Thomas H. McCurdy. Identifying Bull and Bear markets in stock returns. *Journal of Business & Economic Statistics*, 18(1):100–112, January 2000.

34. Kneale T. Marshall and Robert M. Oliver. A constant-work model for student attendance and enrollment. *Operations Research*, pages 193–206, 1970.

35. Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1987.

36. Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley-Interscience, New York, 2000.

37. Prasad A. Naik, Michael R. Hagerty, and Chih-Ling Tsai. A new dimension reduction approach for data-rich marketing environments: Sliced inverse regression. *Journal of Marketing Research*, pages 88–101, 2000.

38. Prasad A. Naik, Peide Shi, and Chih-Ling Tsai. Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102 (477):244–254, 2007.

39. Prasad A. Naik, Michel Wedel, and Wagner Kamakura. Multi-index binary response analysis of large data sets. *Journal of Business & Economics Statistics*, 28(1):67–81, 2010.

40. Oded Netzer, James M. Lattin, and V. Srinivasan. A hidden Markov model of customer relationship dynamics. *Marketing Science*, 27(2):185–204, 2008.

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

42. Peter Rossi and Greg Allenby. Bayesian statistics and marketing. *Marketing Science*, 22 (3):304–328, 2003.

43. Stanley L. Sclove. Population mixture models and clustering algorithms. *Commun. Statist. - Theor. Meth.*, A6(5):417–434, 1977.

44. Stanley L. Sclove. A derivation of the logistic regression model from the model for Gaussian discriminant analysis. Research memorandum, College of Business Administration, University of Illinois at Chicago, 2007.

45. Steven L. Scott, Gareth M. James, and Catherine A. Sugar. Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association*, 100(470): 359–369, 2005.

46. D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

47. Richard D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8:227–245, 1989.

48. A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260C269, 1967.

49. Peiming Wang and Martin L. Puterman. Mixed Logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics*, 3(2):175–200, 2007.

50. J. H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, 1970.

51. Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R.* CRC Press, Boca Raton, FL, 2009.

VITA

EDUCATION

- August 2005-August 2011

  Ph.D., Business Administration with area of inquiry Business Statistics, University of Illinois at Chicago, U.S.A.

- August 2003-May 2005

  Master of Science, Statistical Computing, University of Central Florida, U.S.A.

- August 2000-June 2003

  Master of Science, Applied Mathematics with area of inquiry Accounting Information System, South China University of Technology, China

- August 1996-June 2000

  Bachelor, Accounting, South China University of Technology, China

TEACHING/RESEARCH/WORK EXPERIENCE

- January, 2009-Present

  Project Manger, Marketing Analysis, Discover Financial Services

- January, 2008-December, 2008

  Senior Associate, Marketing Analysis, Discover Financial Services

- June, 2006-December, 2007

  Associate, Marketing Analysis, Discover Financial Services

- August, 2005-May, 2006

  Teaching Assistant for Business Statistics, University of Illinois at Chicago

- January, 2005-June, 2005

  Assistant Consultant, Database Marketing, Hilton Grand Vacation Company

- August 2003-May, 2005

  Teaching Assistant for Statistics, University of Central Florida