Understanding Household's Complex Brand Choice Behaviors:Variety-seeking, Inertia, or Hybrid Behavior?

Department of Social Systems and Management

Discussion Paper Series

No.1300

Understanding Household's Complex Brand Choice Behaviors : Variety-seeking, Inertia, or Hybrid Behavior?

by

Tomohito KAMAI and Yuichiro KANAZAWA

October 2012

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Understanding Household's Complex Brand Choice Behaviors: Variety-seeking, Inertia, or Hybrid Behavior?

Tomohito KAMAI * Yuichiro KANAZAWA †

October 5, 2012

Abstract

For frequently purchased products, household's brand choice behavior may be influenced by its past choices in two ways, namely inertia and variety-seeking. Even though there is vast literature in this field, much of literature assumes either inertia or variety-seeking because these two behaviors were once considered to be mutually exclusive (Givon 1984). One of few studies that entertain the possibility of households transforming itself from initial inertia to variety-seeking later is Bawa (1990), which he calls it hybrid behavior. However, Bawa (1990) does not incorporate the effect of marketing variables and solely relies on run - the number of the purchases of the same

*Doctoral Student, Graduate School of Systems and Information, University of Tsukuba. E-mail: tkamai@sk.tsukuba.ac.jp.

†Professor of Statistics, Department of Social Systems and Management, University of Tsukuba. Email: kanazawa@sk.tsukuba.ac.jp.

brand - to express the hybrid behavior. In this study, the comprehensive model which can accommodate variety-seeking, inertia and hybrid behaviors along with the heterogeneous preferences to brands and sensitivities to marketing variables across households is proposed to express households ’ complex brand choice behaviors. We present the empirical application of our model to a benchmark panel data of ketchup buying households provided by ERIM division of AC Nielsen. The implications of the results and discussions are presented.

1 Introduction

Understanding the dynamic household's brand choice behavior is an interest to both retailers and manufacturers who plan their marketing strategies. One of the major topics in this field includes two behavioral patterns seemingly persisting across more than one purchase occasion. They are referred to as inertia and variety-seeking.

The term "inertia" is defined to be the behavior such that a purchase of a brand will increase the probability of the same brand being purchased again on a succeeding occasion. The "variety-seeking" refers to the opposite, i.e., a purchase of a brand will decrease the probability of the same brand being purchased again on a succeeding occasion (Bawa, 1990; Chintagunta, 1998).[1] Often they are jointly referred to as state dependence.

Accounting for state dependence in household's brand choice behavior is important for many reasons. For example, the ability to measure the household's tendency and intensity of variety-seeking would result in a better un-

[1] When the purchase probability is affected only by the last purchase (i.e., Markov process), the behavior is called "first-order" behavior. Accordingly, "zero-order" indicates the behavior which is independent of any previous choice.

derstanding of the market and would help the firm to improve its competitive position (Givon, 1984). On the other hand, Keane (1997) points out that if a researcher ignores a household's inertial tendency when it exists and only accounts for households-specific preferences to brands, the latter will be exaggerated. Meanwhile, if one ignores the fact that some households purchase a certain brand because they prefer it and only accounts for inertia, the effect of inertia would be overstated.

The existence of state dependence plays important roles in a marketer's decision making as Chintagunta (1998) states: "From a managerial stand point, it would be important to know whether a brand's consumers are inertial or variety prone (Chintagunta, 1998, 254)." For example, when households have inertial tendencies, inducing trials via a free sampling would be an effective promotion. Also, a brand retention could be more emphasized to inertial households as their brand switching could lead to a sustained defection (Chintagunta, 1998). On the other hand, the existence of variety-seeking behavior could motivate marketers to extend their product lines so that household's brand switching behavior would benefit their own products (Seetharaman, 2004).

Moreover, the effect of promotion is more accurately measured by accounting for the multi-periods impacts of promotion due to state dependence.

In this paper, we propose a model to capture complex behaviors of households who vary their choices with or without any apparent reasons. The rest of this paper is structured as follows. In the next section, we review literatures in this field. In section 3, we propose our specifications of utility and model. In section 4, we review the latent class model. In section 5, we will present empirical results and findings. The final section concludes with discussions.

2 Literature Review

The motivation for modeling variety-seeking behavior is well described in the work of McAlister (1982), where she proposes the satiation hypothesis which claims that the preference to a brand decreases when the "inventory of attribute," such as "fruit flavor of soft drink," reaches the point of satiation. Then a variety-seeking behavior is triggered by a desire for different attributes or a stimulus associated with switching behavior. Much of the literature on the variety-seeking is based on this theory (Lattin and McAlister, 1985; Feinberg et al., 1992; Trivedi et al., 1994).

The example of specification of the variety-seeking model can be seen in Trivedi et al. (1994). In their work, they conceptualize the choice being derived from the balance between two forces; the preference for a specific brand (they termed it as a tendency to purchase a brand that is close to household's "ideal") and the propensity to seek variety. The choice probability of brand j given the previous purchase of l in their work is governed by the term

$$u_j(1 - v) + vf(D_{jl})$$

where u_j is the utility of brand j when variety-seeking is absent, D_{jl} is the dissimilarity index between brands j and l ranging between 0 and 1, and v is the degree of variety-seeking sought on a given purchase occasion ranging between 0 and 1. When $v = 0$, the household's choice is not affected by its previous choice (thus becomes zero-order behavior) and when $v = 1$, a household's propensity of variety-seeking is maximum.

The dissimilarity index is used since a variety-seeking effect and the similarities between the two brands are related, i.e., when a household seeks variety, it would choose something "dissimilar" to the previously chosen brand. The higher value of D_{jl} would lead to the higher probability for purchasing j

in such a case. When a household seeks for the similar brand to l, the higher value of D_{jl} would lead to the lower probability for purchasing j.

The idea of incorporating the balance between a genuine preference to a brand and a variety-seeking propensity comes from the work of Givon (1984) and the idea of incorporating the dissimilarity index comes from Lattin and McAlister (1985). The contribution of Trivedi et al. (1994) is the specification of the term v which is assumed to follow a Beta distributions. In the earlier studies, this index was assumed to be fixed.

In contrast to a variety-seeking behavior, some researchers found the inertial tendency for some households. There are many plausible explanations for inertia including a habit persistence, learning effect combined with risk aversion and so forth (Keane, 1997). One of the most influential work in this stream is of Guadagni and Little (1983), which proposes the unique specification of the brand loyalty variable where the weighted sequential influences of past choices to the current utility are incorporated in the form of

$$\begin{aligned} x_{ijt_i} &= \alpha_b \cdot x_{ij,t_i-1} \\ &\qquad +(1-\alpha_b)\{\mathit{household}\ i\ \mathit{bought\ brand}\ j\ \mathit{on\ occasion}\ (t_i-1)\}, \end{aligned}$$

where α_b is a parameter and $\{\mathit{statement}\}$ denotes an indicator function taking an unity if the statement is true. The indices $i = 1, \ldots, N$, $j = 1, \ldots, J$ and $t_i = 1, \ldots, T_i$ denote a household, a brand and purchasing occasion of household i respectively in this study. They find that, based on the criteria such as t-values, the loyalty variable they propose contributes most to the model fit. This may be a part of the reason that their loyalty variable has been widely used in this study field. The other studies on inertia include Gupta et al. (1997), Keane (1997) and Seetharaman (2003).

Meanwhile, Bawa (1990) suggests the "hybrid behavior" which assumes both inertia and variety-seeking behaviors for the same household. The hybrid behavior hypothesized in Bawa (1990) is characterized by the behavior where a household exhibits an inertial tendency for a certain period of time and then exhibits a variety-seeking tendency once a certain period of time passes. In other words, at least for some households, it seems that the marginal utility of the same brand increases first but starts to decrease by the repeated consumption of the same brand, and brand switching occurs once the utility of that brand becomes lower than those of the other brands.[2]

He justifies the hybrid behavior based on the psychological study of Berlyne (1970) which examines a relationship between the stimulus conditions and hedonistic values. According to the study, the hedonistic value such as pleasantness increases at first as a stimulus becomes more familiar, but starts to decrease once the stimulus loses its novelty due to the repeated exposure. This is because any stimulus that produces the moderate increase in arousal is rewarding and pleasant, and familiarity to the stimulus will reduce the arousal level of the aversive stimulus due to its novelty to the point of non-aversive state. Therefore the utility function based on this theory has inverted U-shape when the horizontal axis is taken to represent the consumption period or repeating purchases of the brand.

In the model of Bawa (1990), the marginal utility of the brand is specified to be the function of the number of times the same brand had been continuously purchased up to t_i-th occasion, which he calls "run" and denoted it by $r_{ij}(t_i)$. Accordingly, the utility of household i for brand j on t_i-th occasion

[2]There are some other studies than Bawa (1990) which assume both inertia and variety-seeking behaviors, such as Chintagunta (1998), but these studies usually assume inertia for some households and variety-seeking behaviors for the other households and these behavioral tendencies are assumed to be stable over time.

given $r_{ij}(t_i)$ in Bawa's model is

$$U_{ijt_i|r_{ij}(t_i)} = a_{ij} + b_i r_{ij}(t_i) + c_i(r_{ij}(t_i))^2,$$

where a_{ij} is a brand specific constant, and b_i and c_i are parameters, while the marginal utility of household i for brand l ($\neq j$) up to time t_i conditional on $r_{ij}(t_i)$ is specified as

$$U_{ilt_i|r_{ij}(t_i)} = a_{il},$$

where a_{il} is a brand specific constant for brand l. The terms a_{ij}, a_{il}, b_i and c_i are parameters to be estimated and are assumed to take different values across households.

There are several limitations in the specification of Bawa's model. Besides the model limits the number of brands to be two, it does not incorporate the effects of marketing variables and households' heterogeneous preferences to brands, which are known to be relevant in brand choice behaviors (Keane, 1997).

The major drawback of the model is its specification of an inertial part of the utility; it uses run for the inertial part of the hybrid behavior, which is not generally adopted. The examples of variables used to incorporate inertia include the lagged choices, the lagged utilities, and the serially correlated error structure and so forth (Seetharaman, 2003). Run may be categorized as the lagged choice, but a more general way of including the lagged choice effect is through the loyalty variable as in Guadagni and Little (1983) for instance.

On the other hand, most of the early works on variety-seeking assume the first order Markov process. As stated in the work in Bawa (1990), it may be possible that a variety-seeking propensity emerges in the course of repeated purchases of the same brand, which cannot be formulated by the first order Markov process. To avoid this issue, we have to use a different approach.

In short, the comprehensive model which can accommodate both inertia and variety-seeking behaviors as well as the hybrid behavior along with the heterogeneous preferences to brands and sensitivities to marketing variables is necessary in order to understand the household's complex brand choice behavior accurately. In the next section, we discuss the specification of our model.

3 The Specification of the Model

In constructing the model, we use the brand loyalty variable of Guadagni and Little (1983), which we will refer to it as "GL variable" henceforth, to express inertial part of the hybrid behavior. To capture the effect of variety-seeking, we include run which is defined in Bawa (1990). The purpose of including run is to "put brake" on the utility increase due to the GL variable, which keeps increasing as long as the same brand is kept being purchased. By including run, the utility for the same brand would start to decline as a result of the repeated consumption of the same brand if run negatively affects utility. Another reason for using run is it can express the cumulative effect of satiation due to a repeated consumption of a brand. By this specification, if a household has inertial tendency, the coefficients of both GL variable and run would significantly be non-negative. On the other hand, the variety-seeking behavior could be detected by the non-positive coefficient of run. Meanwhile, the hybrid behavior could be detected by the relative magnitudes of positive coefficient of GL variable and negative coefficient of run.

Now we write the utility of household $i = 1, \ldots, N$ for brand $j = 1, \ldots, J$

on occasion $t_i = 1, \ldots, T_i$ as

$$U_{ijt_i} \quad = \quad \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s + \epsilon_{ijt_i}, \tag{3.1}$$

where $\boldsymbol{x}_{ijt}$ is a $1 \times R$ vector of the explanatory variables including a set of the dummy variables for brands except for one base brand, the shelf price of brand j, a dummy variable for coupon usage times a coupon face value, the dummy variables for feature and display, GL variable and run. The $\boldsymbol{\beta}_s$ is a corresponding $R \times 1$ vector of parameters for segment $s = 1, \ldots, S$. The segment is a subset into which households are placed, where those in the same segment are assumed to be homogeneous in the preferences to brands and responsiveness to the marketing variables. Each household is assumed to belong to only one of the segments, and it is further assumed that household belongs to the same segment over the period of observation. The term ϵ_{ijt_i} is a random error term that captures the effects of unobserved variables which is assumed to follow i.i.d. Gumbel distribution. By this specification, the brand choice probability is given by the multinomial logit model, and the log likelihood function for our the panel data is given by

$$l(\boldsymbol{\beta}) = \sum_{s=1}^{S}\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{h_i(s)\cdot y_{ijt_i}\cdot \ln\left(\frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)}\right)\right\}$$

where y_{ijt_i} is the indicator variable taking an unity if household i purchases a brand j on t_i-th occasion and $h_i(s)$ is the expected value of membership probability of household i to segment s. For an estimation, we use EM algorithm assuming the households' membership probabilities to each segment as missing values.

4 The Latent Class Model

In this study, we propose a model to accommodate the coexistence of households with different behavioral patterns in the same market by using the latent class model. The latent class model is one of the general models to incorporate heterogeneity across households assuming a finite and fixed number of segments. As will be explained, the overall households' choice probabilities are given by the weighted sum of segment-level choice probabilities for the brands in this model (Bucklin et al., 1998). In other words, each of the (overall) unconditional choice probabilities for brands "can be decomposed into a weighted average of underlying (or "latent") choice probabilities (Kamakura and Russell, 1989, 380)."

The idea behind this type of model is that there is an underlying multidimensional distribution of households' heterogeneities (i.e., intrinsic preferences for brands and relative responsiveness to the marketing variables) which characterize their behaviors. In the latent class model, the underlying distribution is assumed to be discrete. Because the finite representations of households' characteristics of latent class model coincide well with the concept of segments in marketing, the model is widely applied to the marketing field. The major work of this field is Kamakura and Russell (1989). The other work using the latent class model include Bucklin et al. (1998) and Gupta and Chintagunta (1994).

4.1 Estimation

Now let us assume there are $s = 1, \ldots, S$ segments in the sample. Obviously as the number of segments is unknown nor can be observed, it must be estimated. First define the relative sizes of segment s as π_s such that

$$0 < \pi_s \leq 1$$

for all s and

$$\sum_{s=1}^{S} \pi_s = 1. \tag{4.1}$$

In the latent class model, each household has different membership probabilities for these segments, because membership probabilities are estimated from their choice histories which differ across households. In fact, the estimate of the relative size of segment π_s is calculated as the mean of the households' membership probabilities for that segment. Accordingly, the term π_s can be viewed as the "likelihood of finding a household in segment s (Kamakura and Russell, 1989, 380)" in the sample. The detail will be explained later.

Now let us define the random variable Y_{ijt_i} which takes value one if household i chooses brand j at t_i-th occasion. In other words, for household i, let y_{ijt_i} be entries of $T_i \times J$ matrix $\boldsymbol{Y}_i$

$$\boldsymbol{Y}_i = \begin{pmatrix} y_{i11}, & \cdots & , y_{iJ1} \\ \vdots & \vdots & \vdots \\ y_{i1T_i}, & \cdots & , y_{iJT_i} \end{pmatrix} \tag{4.2}$$

and let us denote each row as $\boldsymbol{y}_{it_i}$. Since we assume ϵ_{ijt_i}s follow i.i.d. Gumbel distribution, we can express the probability that household i in segment s chooses brand j at the t_i-th occasion in the standard logit form as

$$\Pr\{(y_{i1t_i}, \ldots, y_{iJt_i}) = (\underbrace{0, \ldots, 0}_{j-1}, 1, \underbrace{0, \ldots, 0}_{J-j}) | S_i = s; \boldsymbol{\beta}_s\} = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)}, \tag{4.3}$$

where the random variable S_i indicates which segments household i belongs to, assuming we could observe the segment membership of household i. For brerity, we abbreviate (4.3) as

$$\Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s) = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}_s)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}_s)}, \tag{4.4}$$

henceforth.

The unconditional choice probability for brand j of a randomly selected household i can be obtained by integrating out the equation (4.3) by the density in the population π_s as[3]

$$\Pr(Y_{it_i} = j) = \int \Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s) \cdot \pi_s ds. \tag{4.5}$$

Since the relative size of the segment π_s is discrete, (4.5) is written as

$$\Pr(Y_{it_i} = j) = \sum_{s=1}^{S} \pi_s \cdot \Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s). \tag{4.6}$$

This is a weighted average of logit formula evaluated at each mass point (segment), as pointed out by Kamakura and Russell (1989).

Now suppose that household i has the choice history defined as $H_i = (H_{i1}, \ldots, H_{iT_i})$, where element H_{it_i} indicates the brand purchased at t_i-th occasion. Then the conditional choice probability that household i has the choice history H_i given that household i belongs to segment s is written as

$$\Pr(H_i | S_i = s; \boldsymbol{\beta}_s) = \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} \{\Pr(Y_{it_i} = j | S_i = s; \boldsymbol{\beta}_s)\}^{y_{ijt_i}}. \tag{4.7}$$

In the same manner as (4.6), the unconditional probability of randomly selected household i having the choice history H_i can be written as

$$\Pr(H_i; \boldsymbol{\beta}) = \sum_{s=1}^{S} \pi_s \cdot \Pr(H_i | S_i = s; \boldsymbol{\beta}_s) \tag{4.8}$$

[3]The model of the form (4.5) is sometimes called mixed logit model and π_s is called mixing distribution. The latent class model can be regarded as the special case of mixed logit model where mixing distribution is discrete (Train, 2003).

where $\boldsymbol{\beta}$ is $R \times S$ parameter matrix

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_S) = \begin{pmatrix} \beta_{11}, \ldots, \beta_{1s}, \ldots, \beta_{1S} \\ \vdots \\ \beta_{r1}, \ldots, \beta_{rs}, \ldots, \beta_{rS} \\ \vdots \\ \beta_{R1}, \ldots, \beta_{Rs}, \ldots, \beta_{RS} \end{pmatrix}. \tag{4.9}$$

Here let us define for each i the multinomial indicator random variable $z_i(s)$ which takes one if household i belongs to segment s and 0 otherwise, assuming we know the membership probability of household i belonging to segment s given his/her purchase history H_i, that is $\Pr(S_i = s|H_i; \boldsymbol{\beta}_s)$. Then this membership indicator random variables $z_i(s)$'s are entries of $N \times S$ matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1(\cdot) \\ \vdots \\ \boldsymbol{z}_N(\cdot) \end{pmatrix} = \begin{pmatrix} z_1(1), \ldots, z_1(S) \\ \vdots \\ z_N(1), \ldots, z_N(S) \end{pmatrix}.$$

The row sums of the matrix $\boldsymbol{Z}$ above are all 1.

Assuming we were able to observe $\boldsymbol{Z}$, the likelihood given the choice histories of the all households under consideration is written as[4]

$$L(\boldsymbol{\pi}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z}) = \prod_{i=1}^{N} \prod_{s=1}^{S} \{\pi_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)\}^{z_i(s)},$$

where $\boldsymbol{H} = (H_1, \ldots, H_i, \ldots, H_N)$ is the choice history of all households in the sample, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)$ is $1 \times S$ vector of relative sizes of segments.

[4] The term $\pi_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)$ is the joint probability that household i belongs to segment s and has choice history H_i. Note, however, that the relative size of segment π_s is unknown and has to be estimated.

Accordingly, the log likelihood could be written as

$$\begin{aligned} l(\boldsymbol{\pi}, \boldsymbol{\beta} | \boldsymbol{H}, \boldsymbol{Z}) &= \sum_{i=1}^{N} \sum_{s=1}^{S} z_i(s) \cdot \ln\left(\pi_s \cdot \Pr(H_i | S_i = s; \boldsymbol{\beta}_s)\right) \\ &= \sum_{i=1}^{N} \sum_{s=1}^{S} z_i(s) \cdot \ln \Pr(H_i | S_i = s; \boldsymbol{\beta}_s) + \sum_{i=1}^{N} \sum_{s=1}^{S} z_i(s) \cdot \ln \pi_s. \end{aligned} \tag{4.10}$$

Now if we were able to observe $\boldsymbol{Z}$, the algorithm to estimate parameters $(\boldsymbol{\pi}, \boldsymbol{\beta})$ is as follows.

Step 0.1 Set $t = 0$. Set the initial values $\hat{\boldsymbol{\beta}}_s^{(0)}$ for $s = 1, \ldots, S$ and set $\pi_s^{(0)} = 1/S$ for $s = 1, \ldots, S$.

Step 0.2 Calculate $l^{(t)}\left(\boldsymbol{\pi}^{(t)}, \hat{\boldsymbol{\beta}}^{(t)} | \mathbf{H}, \mathbf{Z}\right)$ using (4.10).

Step 1 Calculate $\pi_s^{(t+1)}$ for $s = 1, \ldots, S$ from the method which will be explained below.

Step 2 Estimate $\hat{\boldsymbol{\beta}}_s^{(t+1)}$ for $s = 1, \ldots, S$ using the scoring or Newton-Raphson method.

Step 3 Calculate $l^{(t+1)}\left(\boldsymbol{\pi}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)} | \mathbf{H}, \mathbf{Z}\right)$ using (4.10). If $l^{(t+1)}\left(\boldsymbol{\pi}^{(t+1)}, \hat{\boldsymbol{\beta}}_s^{(t+1)} | \mathbf{H}, \mathbf{Z}\right)$ and $l^{(t)}\left(\boldsymbol{\pi}^{(t)}, \hat{\boldsymbol{\beta}}_s^{(t)} | \mathbf{H}, \mathbf{Z}\right)$ are close enough, for example, less than some small prescribed constant ϵ, stop the iteration as the likelihood is maximized. Else set $t = t + 1$ and goto **Step 1**.

However, in reality, we cannot possibly obtain the information $z_i(s)$. In such a situation, the method called EM algorithm may be implemented to obtain the estimate of $z_i(s)$ along with the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ as explained in the following subsection.

4.2 EM algorithm

If the segment memberships of households $\mathbf{Z}$ were completely known, the vector of parameters $\boldsymbol{\beta}_s$ can be estimated by the algorithm described above using well-known methods such as Newton-Raphson method. EM algorithm takes advantage of this fact and in the algorithm, the household's membership to the segment $z_i(s)$ is assumed at first to be missing values and this value is imputed by its "expectation" (to be explained below). Then the conditional likelihood is maximized based on the expected value of membership to the segment. The households' expected membership is then updated using the updated likelihood. This cycle of "expectation" of membership to the segment and "maximization" of likelihood is repeated until the likelihood converges.

Now taking the expectation with respect to $z_i(s)$ for the log likelihood (4.10), we have

$$E[l(\boldsymbol{\pi}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z})] = \sum_{i=1}^{N}\sum_{s=1}^{S} h_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} h_i(s) \cdot \ln \pi_s, \tag{4.11}$$

where

$$h_i(s) = E[z_i(s)] = \sum_{l=1}^{S} z_i(l) \cdot \Pr(S_i = l|H_i; \boldsymbol{\beta}_l) = \Pr(S_i = s|H_i; \boldsymbol{\beta}_s) \tag{4.12}$$

is the expected values of the indicator random variable $z_i(s)$ for $s = 1, \ldots, S$. Since parameter $\boldsymbol{\beta}$ in (4.9) only appears in the first term and $\boldsymbol{\pi}$ only appears in the second term on the right hand side of (4.11), they can be estimated by maximizing $E[l(\boldsymbol{\pi}, \boldsymbol{\beta}|\boldsymbol{H}, \boldsymbol{Z})]$ alternately.

Let us first look at the second term on the right hand side of (4.11). Since we have the condition $\sum_{s=1}^{S} \pi_s = 1$ from (4.1), the second term can be maximized by the method of Lagrange multipliers given $\boldsymbol{\beta}_s$. Set

$$L = \sum_{i=1}^{N} \sum_{s=1}^{S} h_i(s) \cdot \ln \pi_s - \lambda \left\{ \sum_{s=1}^{S} \pi_s - 1 \right\}.$$

Then we have $(S+1)$ set of equations by partially differentiating L with respect to π_s's and λ. Now we set resulting formulas zero as

$$\begin{cases} \frac{\partial L}{\partial \pi_1} = \frac{\sum_{i=1}^{N} h_i(1)}{\pi_1} - \lambda = 0, \\ \qquad \vdots \\ \frac{\partial L}{\partial \pi_S} = \frac{\sum_{i=1}^{N} h_i(S)}{\pi_S} - \lambda = 0, \\ \frac{\partial L}{\partial \lambda} = -\sum_{s=1}^{S} \pi_s + 1 = 0. \end{cases} \tag{4.13}$$

From the first S equations in (4.13), we have

$$\lambda = \frac{\sum_{i=1}^{N} h_i(s)}{\pi_s}, \tag{4.14}$$

or

$$\pi_s = \frac{1}{\lambda} \sum_{i=1}^{N} h_i(s) \tag{4.15}$$

for $s = 1, \ldots, S$. Substitute these equations into the last equation in (4.13) to obtain

$$\frac{1}{\lambda} \sum_{i=1}^{N} h_i(1) + \cdots + \frac{1}{\lambda} \sum_{i=1}^{N} h_i(S) = 1$$

or

$$\sum_{i=1}^{N} (h_i(1) + \cdots + h_i(S)) = \lambda$$

or

$$N = \lambda,$$

since $h_i(1) + \cdots + h_i(S) = 1$. Therefore we have from (4.15)

$$\pi_s = \frac{\sum_{i=1}^{N} h_i(s)}{N} \tag{4.16}$$

for $s = 1, \ldots, S$. The solution (4.16) means that the relative size of segment s is the average of segment membership for s across all households in the sample.

Now from (4.12), $h_i(s) = \Pr(S_i = s|H_i; \boldsymbol{\beta}_s)$ can be calculated using the definition of conditional probability as[5]

$$h_i(s) = \frac{\Pr(S_i = s, H_i; \boldsymbol{\beta}_s)}{\Pr(H_i; \boldsymbol{\beta})} = \frac{\pi_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)}{\sum_{s=1}^{S} \pi_s \cdot \Pr(H_i|S_i = s; \boldsymbol{\beta}_s)}. \tag{4.17}$$

By substituting (4.17) for (4.16), we obtain π_s.

As for the first term of the right hand side of (4.11) for segment s, the parameters can be estimated independently for each segment since the vectors of parameters $\boldsymbol{\beta}_s$ are independent across segments. Then the first term on the right hand side of (4.11) for segment s is written with the notation similar to (4.7) as[6]

$$\begin{aligned} l_s(\boldsymbol{\beta}_s|\boldsymbol{H}) &= \sum_{i=1}^{N} h_i(s) \cdot \ln \Pr(H_i|S_i = s; \boldsymbol{\beta}_s) \\ &= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \{h_i(s) \cdot y_{ijt_i} \cdot \ln \Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)\}. \end{aligned} \tag{4.18}$$

To implement EM algorithm, repeat the following steps.

EM algorithm

[5] Note that $h_i(s)$ in (4.17) can be interpreted as the posterior distribution of household i's membership probability for segment s with prior distribution π_s and likelihood H_i given segment membership $S_i = s$ as we mentioned earlier.

[6] For more detail, see Appendix.

Step 0.1 Set $t = 0$. Set the initial values $\hat{\boldsymbol{\beta}}_s^{(0)}$ for $s = 1, \ldots, S$ and set $\pi_s^{(0)} = 1/S$ for $s = 1, \ldots, S$.

Step 0.2 Set $s = 1$. For $i = 1, \ldots, N$, calculate $h_i^{(t)}(s)$ by calculating $\Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)$ using (4.3) first then (4.7) and (4.8) successively with $\hat{\boldsymbol{\beta}}_s^{(t)}$ and $\pi_s^{(t)}$ and substitute these interim results for (4.17). Set $s = s + 1$ and repeat **Step 0.2** until $s = S$.

Step 0.3 Calculate $E\left[l^{(t)}\left(\boldsymbol{\pi}^{(t)}, \hat{\boldsymbol{\beta}}^{(t)}|\mathbf{H}, \mathbf{Z}\right)\right]$ using (4.11).

Step 1 Set $s = 1$. Renew $\pi_s^{(t+1)}$ from (4.16) using $h_i^{(t)}(s)$.

Step 2 Estimate $\hat{\boldsymbol{\beta}}_s^{(t+1)}$ by maximizing (4.18) with (4.3) and $h_i^{(t)}(s)$ obtained previously. The actual maximization is done by the scoring or Newton-Raphson method.

Step 3 Renew $\Pr(Y_{it_i} = j|S_i = s; \boldsymbol{\beta}_s)^{(t+1)}$ by substituting $\hat{\boldsymbol{\beta}}_s^{(t+1)}$ obtained in **Step 2**.

Step 4 Calculate $h_i^{(t+1)}(s)$ from (4.17) with the renewed $\hat{\boldsymbol{\beta}}_s^{(t+1)}$ and $\pi_s^{(t+1)}$ for $i = 1, \ldots, N$. Set $s = s + 1$ and goto **Step 1**. If $s = S$, goto **Step 5**.

Step 5 Calculate $E\left[l^{(t+1)}\left(\boldsymbol{\pi}^{(t+1)}, \hat{\boldsymbol{\beta}}^{(t+1)}|\mathbf{H}, \mathbf{Z}\right)\right]$ using (4.11). If $E\left[l^{(t+1)}\left(\boldsymbol{\pi}^{(t+1)}, \hat{\boldsymbol{\beta}}_s^{(t+1)}|\mathbf{H}, \mathbf{Z}\right)\right]$ and $E\left[l^{(t)}\left(\boldsymbol{\pi}^{(t)}, \hat{\boldsymbol{\beta}}_s^{(t)}|\mathbf{H}, \mathbf{Z}\right)\right]$ are close enough, for example, less than small prescribed constant ϵ, stop the iteration as the expected log likelihood is maximized. Else set $s = 1$ and $t = t+1$, and return to **Step 1**.

5 Empirical Results

We use ERIM database, the panel data of U.S. households in Sioux Falls, SD and Springfield, MO which was collected from 1st week of 1986 to 34th week of 1988. ERIM database is the data collected by the now-defunct ERIM division of A.C. Nielsen on panels of households in Sioux Falls and Springfield for academic research.[7]

We choose a ketchup category for our empirical analysis for the following reasons. First, because we are interested in a household's brand choice behavior with the possible presence of state dependence, product categories in which a household exhibits a strong genuine preference to specific brand are not suitable because a household would buy the specific brand anyway. Secondly, the products that are purchased with relatively high frequency are preferable, since we incorporate the effect of the past brand choice on the current purchasing occasion. In other words, the products that are purchased on irregular intervals would not suit our purpose as households may forget the brands they purchased on the previous occasion.

Now we build two sets of panel data from the ERIM database; one for Sioux Falls and the other for Springfield. Sioux Falls data consists of 1,693 households who purchased ketchup during the data collection period, while Springfield data consists of 1,343 households. Separate data sets are set up for these two markets because households may behave differently between these two markets, but Springfield data turns out to have a very limited

[7]We acknowledge the James M. Kilts Center, University of Chicago Booth School of Business for letting us use the data.

number of households after data screening.

Because there are more than forty Stock Keeping Units (SKUs) in the original panel data, we use the following criteria to select SKUs for our analysis: First, we drop the SKUs whose market shares are less than 1% because some households in these data sets are not able to choose them because too few stores in each of the markets carry them. This leaves fourteen SKUs for Sioux Falls.[8] Next, we choose the SKUs whose sizes are either 32 or 28 ounce, which seem to be the standard sizes of ketchup judged by their market shares; they account for 81.6% and 80.0% of the market shares in Sioux Falls and Springfield respectively in the panel data. The other sizes include 14, 40, 44 and 64 ounce, but those who buy the ketchup of these sizes may have different demographic characteristics and thus may have different purchasing patterns from those who buy the standard sized ketchup. This leaves eight and eleven SKUs in respective markets. There are 516 households who choose ketchup from the eight SKUs with 3,933 purchase records in Sioux Falls.[9]

Next, we check how many stores carry all these SKUs in the respective market, because if households buy ketchup from the other stores than those carrying all SKUs, their SKU selections could have been influenced by the lack of selection. We first take a look at Sioux Falls. There are fifteen stores in Sioux Falls, but only five of them carry all eight SKUs.[10] If we removed the households that bought at least one ketchup in stores other than these five, only 120 households with 497 purchase records would be left for the analysis.

[8]Eighteen SKUs remain for Springfield at this stage.

[9]There are 494 households who buy ketchup from the eleven SKUs with 3,026 purchase records in Springfield.

[10]We assume that the store carries the SKU if at least one purchase record of the SKU is found in that store during the data collection period.

The large reduction of data is because the sixth, seventh, and eighth selling SKUs are simultaneously available only in few stores in Sioux Falls. Hence we choose to retain top selling five SKUs. Among fifteen stores in Sioux Falls, twelve of them carry all top selling five SKUs. After eliminating the households who purchased ketchup in stores other than these twelve, 255 households with 1,791 purchase records remain.

Finally, since we are interested in the household's brand choice behavior across time, we choose to retain the households who made more than or equal to five purchases of ketchup during the period, which leaves 137 households with 1,504 purchase records.[11] After screening data, we collect household ID, SKU purchased, its shelf price, coupon values (when used), store ID, date of purchase, an indicator variable whether it was displayed and an indicator variable whether it was featured on each purchasing occasion.

Since the price of a particular SKU is available only when it is actually purchased, we need to estimate purchase prices (i.e. the price households would have paid if they had coupons) of all the competing SKUs in calculating utilities every time one of the households under study purchases ketchup from the five SKUs. Unfortunately however, the information of the availability of coupons for the SKUs that were not purchased by household i is not available in general. Hence we assume that coupons for the not-purchased-SKUs were

[11] By the same criteria, only 14 households with 109 purchase records remain for Springfield. One of the major differences in these two markets is the number of the stores that carry top selling five SKUs; in Sioux Falls, twelve out of fifteen stores carry them, while only nine out of twenty-one stores do so in Springfield. This difference coupled with the smaller number of households in Springfield relative to Sioux Falls (0.8 to 1) and the difference in market shares of top selling five SKUs between two markets (77.8% of share in Sioux Falls relative to 65.8% of share in Springfield) resulted in the very small number of households in Springfield with only 109 purchase records after data screening.

not available to that household. Now in order to estimate the shelf prices of competing SKUs, we follow the algorithm as follows: First, for a particular store and a particular day, if the other household purchased a competing brand, we assume the shelf price, not the purchasing price, for this SKU is the price household i faced in that store on that day. If no purchase record was found by that criterion, we look for the prices for other stores on that day. If purchase record was still not found, we used the prices in the following dates and weeks.

The summary statistics of the five SKUs analyzed in this study are listed in Table 5.1. "Coupon Usage" indicates the percentage of purchases with coupons, "Display" and "Feature" indicate the percentage of purchases with these promotions, and "Mean Value of Coupons" indicates the average value of coupons used.

Table 5.1: Summary statistics of SKUs under study.

SKU	Share	Mean Price per oz.	Mean Value of Coupons	Coupon Usage	Display	Feature
Heinz 32 oz.	31.70%	3.37	1.24	37.94%	11.52%	43.82%
Heinz PLS 28 oz.	15.80%	4.38	2.41	33.09%	16.73%	34.55%
Hunt's PLS & GLS 32 oz.	14.30%	3.22	1.30	32.57%	11.93%	36.70%
Del Monte 32 oz.	6.40%	2.87	1.00	7.20%	11.20%	36.00%
Control 32 oz.	5.00%	2.65	1.64	3.77%	5.66%	24.53%

To calibrate the effectiveness of our model, we tested the two other models; the model which only uses marketing variables as explanatory variable (Model 1); the model which incorporates GL variable along with the marketing variables (Model 2). The third model is our proposal model which incorporates GL variable and run in addition to marketing variables (Model 3).

We have determined the number of segments based on AIC. The number of segments is chosen to be four because no significant increase in AIC is ob-

Table 5.2: AIC of the three models with different numbers of segments.

	Model 1	Model 2	Model 3
2 segments	1310.60	1058.19	1041.80
3 segments	948.52	861.27	835.64
4 segments	842.64	805.61	796.74
5 segments	819.09	796.54	794.21
6 segments	809.53	803.84	813.02

served for Model 3 when the number of segments is increased from four to five as shown in Table 5.2. The estimated parameters of Model 3 are presented in Table 5.3. The coefficients of the SKU indicate intrinsic preferences for that SKU with respect to "Control 32 ounce" which we choose as the base SKU.

All the coefficients of Model 3 are consistent with the expected economic behaviors, i.e., all the coefficients of price are negative, those of coupon, display and feature are all positive in all segments. However, the intrinsic preferences to SKUs and responsiveness to marketing variables differ significantly across segments. As for GL variable and run, they show interesting patterns which would have not been discovered if run was not included in the model.

Now to see the behavioral patterns regulated by the model, we calculated the purchasing probabilities for each SKU and segment, assuming the situation where a household repeatedly purchase the same brands five times in row. In the calculation, we used the average prices of SKUs and assuming no promotions took place during the period. The results are shown in Table 5.4. For example, the number at $t = 3$ is the purchasing probability of the SKU given two consecutive purchases of that SKU.

Here we summarize the result in short. As shown in Table 5.4, segments 1

Table 5.3: The parameters of the Model 3.

	segment 1	segment 2	segment 3	segment 4
Heinz 32 oz.	-0.076	4.279	1.885	1.437
	(0.0148)	(0.0183)	(0.0122)	(0.0155)
Heinz PLS 28 oz.	1.311	2.661	1.727	2.399
	(0.0072)	(0.0070)	(0.0065)	(0.0073)
Hunt's PLS & GLS 32 oz.	0.027	0.296	2.713	-2.530
	(0.0068)	(0.0050)	(0.0097)	(0.0055)
Del Monte 32 oz.	1.176	-2.141	1.071	-1.154
	(0.0071)	(0.0014)	(0.0043)	(0.0057)
Price	-0.848	-0.735	-0.666	-2.509
	(0.0701)	(0.0769)	(0.0682)	(0.0713)
Coupon	2.792	5.015	3.287	5.471
	(0.0211)	(0.0207)	(0.0183)	(0.0202)
Display	3.419	3.903	3.855	4.750
	(0.0070)	(0.0067)	(0.0067)	(0.0072)
Feature	5.622	2.504	2.938	5.894
	(0.0126)	(0.0121)	(0.0119)	(0.0127)
GL	4.303	0.743	1.737	5.607
	(0.0136)	(0.0115)	(0.0102)	(0.0126)
Run	0.505	*-0.127	-0.222	-0.123
	(0.0891)	(0.0663)	(0.0519)	(0.0529)
Size of Segments	0.261	0.234	0.289	0.217
Total Log Likelihood	-365.81			

* 90% level significance with t-value -1.917. All the other coefficients were significant at the 0.05 level. The numbers in parentheses are standard errors.

Table 5.4: Logit probability of purchase: Model 3

Heinz 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	8.8%	87.6%	20.6%	33.5%
t=1	56.1%	91.7%	42.4%	93.6%
t=2	74.9%	91.1%	40.4%	95.3%
t=3	86.7%	90.5%	37.7%	96.2%
t=4	93.1%	89.7%	34.7%	96.8%
t=5	96.4%	88.8%	31.3%	97.1%

Heinz PLS 28 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	15.0%	8.3%	9.0%	7.1%
t=1	70.1%	12.4%	21.9%	68.7%
t=2	84.5%	11.7%	20.5%	75.3%
t=3	92.3%	10.9%	18.8%	79.4%
t=4	96.1%	10.0%	16.8%	81.9%
t=5	98.0%	9.2%	14.8%	83.5%

Hunt's PLS & GLS 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	11.2%	1.8%	52.5%	0.9%
t=1	62.4%	2.8%	75.8%	21.6%
t=2	79.5%	2.7%	74.2%	27.6%
t=3	89.4%	2.5%	72.0%	32.6%
t=4	94.6%	2.3%	69.3%	36.3%
t=5	97.2%	2.0%	66.0%	38.8%

Del Monte 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	47.4%	0.2%	12.8%	9.0%
t=1	92.2%	0.3%	29.4%	74.0%
t=2	96.5%	0.3%	27.7%	79.8%
t=3	98.4%	0.3%	25.5%	83.3%
t=4	99.2%	0.3%	23.1%	85.5%
t=5	99.6%	0.2%	20.5%	86.7%

Control 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	17.6%	2.1%	5.1%	49.5%
t=1	73.9%	3.2%	13.2%	96.6%
t=2	86.9%	3.0%	12.3%	97.5%
t=3	93.5%	2.8%	11.1%	98.0%
t=4	96.8%	2.5%	9.9%	98.3%
t=5	98.3%	2.3%	8.6%	98.5%

Table 5.5: Logit probability of purchase: Model 2

Heinz 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	8.8%	87.6%	20.6%	33.5%
t=1	94.6%	63.4%	85.1%	32.2%
t=2	95.5%	68.4%	87.3%	35.3%
t=3	96.2%	72.0%	88.9%	38.0%
t=4	96.7%	74.8%	90.0%	40.1%
t=5	97.0%	76.8%	90.8%	41.9%

Heinz PLS 28 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	15.0%	8.3%	9.0%	7.1%
t=1	14.0%	49.7%	70.8%	14.4%
t=2	16.7%	55.2%	74.5%	16.2%
t=3	19.1%	59.5%	77.3%	17.8%
t=4	21.2%	62.8%	79.3%	19.1%
t=5	23.1%	65.4%	80.8%	20.3%

Hunt's PLS & GLS 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	11.2%	1.8%	52.5%	0.9%
t=1	12.4%	77.4%	0.1%	11.5%
t=2	14.9%	81.0%	0.2%	13.0%
t=3	17.1%	83.6%	0.2%	14.3%
t=4	19.1%	85.4%	0.2%	15.5%
t=5	20.8%	86.8%	0.3%	16.5%

Del Monte 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	47.4%	0.2%	12.8%	9.0%
t=1	32.0%	0.8%	9.5%	39.5%
t=2	36.7%	0.9%	11.2%	42.9%
t=3	40.7%	1.1%	12.8%	45.7%
t=4	43.9%	1.3%	14.2%	47.9%
t=5	46.5%	1.4%	15.4%	49.7%

Control 32 oz.	Segment 1	Segment 2	Segment 3	Segment 4
t=0	17.6%	2.1%	5.1%	49.5%
t=1	23.9%	55.4%	6.3%	79.6%
t=2	27.8%	60.7%	7.5%	81.8%
t=3	31.3%	64.9%	8.7%	83.4%
t=4	34.2%	68.0%	9.7%	84.6%
t=5	36.7%	70.4%	10.5%	85.5%

and 4 exhibit strong inertia while segments 2 and 3 exhibit weak and modest variety-seeking tendencies respectively. Segment 2 can be characterized by its strong preferences to Heinz SKUs. Segment 3 with the largest proportion of the households is least price sensitive and has relatively low coefficients for coupons and features. Consumers in segment 4 most price sensitive and respond to promotions most.

6 Conclusions and Discussions

In this study, we develop the comprehensive model which can accommodate inertia and/or variety-seeking dependence along with the heterogeneous preferences to brands and sensitivities to marketing variables and applied it to the panel data of ketchup. Our model achieves the best AIC compared to the previously proposed models with a fair number of significant variables, indicating that the households are heterogeneous in their behavioral patterns over time. The heterogeneities across households in terms of intrinsic brand preferences and sensitivities to marketing variables are modeled as well through the use of the latent class model.

Overall, the results of the empirical analysis give important connotations for marketers, because the model without state dependence could be misleading in constructing the strategy and planning promotions (Keane, 1997). Specifically, the information in Table 5.4 can be used as a starting point for brand managers to plan their marketing strategies and promotional activities. For example, since households in segment 1 exhibit strong inertia, Del Monte may need to ensure it has enough amounts of promotions to retain households in this segment since households in segment 1 are the best customers for Del Monte. As households in this segment have low coefficients for coupon and display but have high coefficient for feature, Del Monte may increase a budget for features to retain households in segment 1. Since households in segment 2 exhibit strong preferences to Heinz, there seems to be little chance for the other brands to be selected. From Heinz perspective, it may not need spending on promotions for this segment since households in this segment buy the brand anyway. Segment 3 is the main target for Hunt's. Because households in this segment are least price sensitive and have relatively low coefficients for coupons and features, the display may be the best way to ensure households in this segment to buy Hunt's. Since households in this segment are variety-seeker, Hunt's may consider the brand line extension. Segment 4 also exhibits strong inertia but they are most price sensitive and respond to promotions most. For Heinz, this is the segment worth promoting its Heinz 32 ounce. The Hunt's and Del Monte are better off to spend its budget on segment 1 or 3 since they have little chance to attract households in segment 4.

Also, the results presented in Table 5.4 along with that of Table 5.3 can also be used to measure the effect of promotion incorporating the multi-period impact by examining the repeat purchasing probabilities. Let us take Heinz 32 ounce as an example. The purchase probabilities for the SKU with and without feature at $t = 0$ are shown in Table 6.1.

Now let M be the size of the market and PM be the profit margin of the SKU. Then the profit for Heinz 32 ounce without any promotion is "$33.0\% \cdot M \cdot PM$" where 33.0% is the proportion of households who purchase the SKU in that condition. The profit when feature takes place is "$93.7\% \cdot M \cdot PM - (\textit{Cost of Feature})$" given that 93.7 % of households buy the SKU given the promotion. Then the incremental profit from promotion is: "$(93.7\% - 33.0\%) \cdot M \cdot PM - (\textit{Cost of Feature})$." The measurement of promotion effect

Table 6.1: An effect of promotion: a feature on Heinz 32 oz.

Segment	1	2	3	4	Average
Segment sizes	26.1%	23.4%	28.9%	21.7%	
without feature	9.4%	81.1%	21.1%	25.3%	33.0%
with feature	96.6%	98.1%	83.5%	99.2%	93.7%
t=1	56.1%	91.7%	42.4%	93.6%	*65.6%
t=2	74.9%	91.1%	40.4%	95.3%	53.1%
t=3	86.7%	90.5%	37.7%	96.2%	46.6%
t=4	93.1%	89.7%	34.7%	96.8%	42.5%
t=5	96.4%	88.8%	31.3%	97.1%	39.6%

* It is the weighted average of the households who would buy the SKU at $t = 1$, i.e., $0.261 \cdot 0.966 \cdot 0.561 + \cdots + 0.217 \cdot 0.992 \cdot 0.936 = 0.656$.

without inertia stops here. However, there would be incremental households who would keep buying Heinz 32 ounce from the next purchasing occasion on due to inertia. According to the Table 6.1, 56.1% of households in segment 1 who bought Heinz 32 ounce would choose the SKU again on the next purchasing occasion assuming no promotions of competing SKUs take place and price stays the same. On the next purchasing occasion, 74.9% among them would buy the SKU again. As such, we can calculate the proportion of the households who would choose Heinz 32 ounce on the succeeding occasions for each segment. The number of households who purchase the SKU are 65.6%, 53.1%, 46.6%, 42.5% and 39.6% times M. Thus the incremental profit up to $t = 5$ can be calculated as

$$M \{(65.6\% - 33.0\%) \cdot PM + \cdots + (39.6\% - 33.0\%) \cdot PM\} - (\mathit{Cost\ of\ Feature})$$
$$= 0.8225 \cdot M \cdot PM - (\mathit{Cost\ of\ Feature}).$$

Thus, the effect of promotion is under-estimated if the effect of inertia is not accounted for. This number would be misleadingly larger if we use Model 2.

Unfortunately, the hybrid behavior was not detected in our analysis. This may be because most of consecutive purchases of the same SKU are three at most; about 90% of purchases in the data are shorter than three runs. Also, the product like ketchup, where the bottle is consumed through a relatively long period of time, a satiation effect may start during the consumption period and that may lead households to switch brand, i.e., the hybrid behavior is hidden as a result of the large package sizes of the ketchup. By using the products which are consumed in a relatively short period of time the hybrid behavior may has been detected.

For future studies, the model presented in this study can be tested using different data sets for the validity of the model. The new variable to explain state dependence can be constructed as well. It would be nice if we can incorporate the budget constraint in the model because households may switch brands depending on their budget at each shopping trip.

Appendix

A Scoring (Newton-Raphson) method

In order to maximize the (log) likelihood function, the algorithm called scoring or Newton-Raphson method is widely used. First we explain Newton-Raphson method. We denote the $\boldsymbol{\beta}$ at $t+1$-th iteration by adding superscript as $\boldsymbol{\beta}^{(t+1)}$. In this method, a second-order Taylor expansion of $LL(\boldsymbol{\beta}^{(t+1)})$ around $LL(\boldsymbol{\beta}^{(t)})$ is taken as

$$LL(\boldsymbol{\beta}^{(t+1)}) = \left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right)^T g_t + \frac{1}{2}\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right)^T H_t \left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right), \tag{A.1}$$

where $R \times 1$ vector g_t is the gradient at $\boldsymbol{\beta}^{(t)}$

$$g_t = \left(\frac{\partial LL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$$

and $R \times R$ matrix H_t is matrix of the second derivatives

$$H_t = \left(\frac{\partial g_t}{\partial \boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \left(\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(t+1)}$.

The value of $\boldsymbol{\beta}^{(t+1)}$ maximizing (A.1) is obtained by setting its derivative zero as

$$\frac{\partial LL\left(\boldsymbol{\beta}^{(t+1)}\right)}{\partial \boldsymbol{\beta}^{(t+1)}} = g_t + H_t\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right) = \mathbf{0}.$$

This means

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-H_t^{-1} g_t\right)$$

assuming H_t is invertible.

In implementing the algorithm, the step size is often adjusted through parameter $\lambda = 0.5, 0.25$ etc as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \lambda \left(-H_t^{-1} g_t\right).$$

It is because as the likelihood function gets close to its maximum, $\boldsymbol{\beta}^{(t+1)}$ may pass the maximal point if its step size $\left(-H_t^{-1} g_t\right)$ is too large.

The scoring method is version of Newton-Raphson method where by the likelihood function $l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)$ is replaced with its expected value $\mathrm{E}\left[l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)\right]$ to reduce the average number of iterations that can fluctuate from sample to sample if we employ the random $l_s^{(t)}\left(\boldsymbol{\pi}, \boldsymbol{\beta}_s^{(t)} | \boldsymbol{H}, \boldsymbol{Z}\right)$.

A.1 Estimating parameters

In this subsection, we demonstrate how to calculate gradient and Hessian using the standard logit specification. We assume that a panel data of purchase histories for consumers $i = 1, \ldots, N$ who purchase one of $j = 1, \ldots, J$ products at $t_i = 1, \ldots, T_i$ occasions. We also assume that all the products are available for the group of consumers

The standard logit model, by modifying (4.3), can be written as

$$\Pr(j|\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}$$

where $\boldsymbol{x}_{ijt_i}$ is $1 \times R$ vector and $\boldsymbol{\beta}$ is $R \times 1$ parameter vector. Accordingly the likelihood function of the logit model can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{t_i=1}^{T_i} \prod_{j=1}^{J} (\Pr(j|\boldsymbol{\beta}))^{y_{ijt_i}},$$

and the log likelihood is written as

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} y_{ijt_i}\log(\Pr(j|\boldsymbol{\beta})) \\ &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} y_{ijt_i}\log(\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) - \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})) \\ &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} y_{ijt_i}\left\{\boldsymbol{x}_{ijt_i}\boldsymbol{\beta} - \log\left(\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\right)\right\}. \end{aligned} \tag{A.2}$$

The gradient Now differentiate (A.2) with respect to the vector $\boldsymbol{\beta}$, we have tentatively

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} y_{ijt_i}\left\{\boldsymbol{x}_{ijt_i}^{T} - \frac{\left(\frac{\partial \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}\right\}, \tag{A.3}$$

since

$$\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_R} \end{bmatrix} = \begin{bmatrix} x_{ijt_i1} \\ \vdots \\ x_{ijt_iR} \end{bmatrix} = \boldsymbol{x}_{ijt_i}^{T}.$$

The last term on the right hand side of (A.3) is

$$\begin{aligned} \frac{\partial \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{l=1}^{J}\begin{bmatrix} \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_R} \end{bmatrix} = \sum_{l=1}^{J}\begin{bmatrix} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})x_{ilt_i1} \\ \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})x_{ilt_i2} \\ \vdots \\ \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})x_{ilt_iR} \end{bmatrix} \\ &= \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\begin{bmatrix} x_{ilt_i1} \\ x_{ilt_i2} \\ \vdots \\ x_{ilt_iR} \end{bmatrix} = \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\boldsymbol{x}_{ilt_i}^{T}. \end{aligned} \tag{A.4}$$

Substituting (A.1) back into (A.3) yields

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J} y_{ijt_i}\left\{\boldsymbol{x}_{ijt_i}^T - \frac{\sum_{l=1}^{J}(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\boldsymbol{x}_{ilt_i}^T)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}\right\} \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{y_{ijt_i}\boldsymbol{x}_{ijt_i}^T - \frac{y_{ijt_i}\sum_{l=1}^{J}\left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\boldsymbol{x}_{ilt_i}^T\right)}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}\right\} \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{y_{ijt_i}\boldsymbol{x}_{ijt_i}^T - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}^T}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}\right\} \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{y_{ijt_i}\boldsymbol{x}_{ijt_i}^T - \Pr_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}^T\right\} \\
&= \sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{y_{ijt_i} - \Pr_i(j|\boldsymbol{\beta})\right\}\boldsymbol{x}_{ijt_i}^T. \qquad (A.5)
\end{aligned}$$

This is $R \times 1$ vector of gradient.

Hessian Now differentiate (A.5) further with respect to $\boldsymbol{\beta}^T$ to obtain tentatively

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= \frac{\partial\left(\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\left\{y_{ijt_i} - \Pr_i(j|\boldsymbol{\beta})\right\}\boldsymbol{x}_{ijt_i}^T\right)}{\partial \boldsymbol{\beta}^T} \\
&= -\sum_{i=1}^{N}\sum_{t_i=1}^{T_i}\sum_{j=1}^{J}\frac{\partial \Pr_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\boldsymbol{x}_{ijt_i}^T, \qquad (A.6)
\end{aligned}$$

where from

$$\frac{\partial \Pr_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = \frac{\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) - \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})\frac{\partial \sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}}{(\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}))^2}. \qquad (A.7)$$

Since the term $\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T$ becomes

$$\begin{aligned}
\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \left[\frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\partial \beta_R}\right] \\
&= [\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})x_{ijt_i1}, \ldots, \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})x_{ijt_iR}] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})[x_{ijt_i1}, \ldots, x_{ijt_iR}] = \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})\boldsymbol{x}_{ijt_i},
\end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}} \cdot \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}^T} \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot \left[\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_R}\right] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot [x_{ijt_i1}, \ldots, x_{ijt_iR}] \\
&= \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \cdot \boldsymbol{x}_{ijt_i},
\end{aligned}
$$

the equation (A.7) becomes

$$
\begin{aligned}
\frac{\partial \mathrm{Pr}_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} &= \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})\boldsymbol{x}_{ijt_i} \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) - \exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta}) \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}}{(\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}))^2} \\
&= \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})\boldsymbol{x}_{ijt_i}}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \boldsymbol{x}_{ijt_i} \\
&= \mathrm{Pr}_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i} - \{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2 \boldsymbol{x}_{ijt_i}. \qquad \text{(A.8)}
\end{aligned}
$$

Substituting (A.8) back into (A.6) yields

$$
\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= -\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \frac{\partial \mathrm{Pr}_i(j|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \boldsymbol{x}_{ijt_i}^T \\
&= -\sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left(\mathrm{Pr}_i(j|\boldsymbol{\beta})\boldsymbol{x}_{ijt_i} - \{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2 \boldsymbol{x}_{ijt_i}\right) \boldsymbol{x}_{ijt_i}^T \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \left(\{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2 \boldsymbol{x}_{ijt_i}^T \boldsymbol{x}_{ijt_i} - \mathrm{Pr}_i(j|\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}^T \boldsymbol{x}_{ijt_i}\right) \\
&= \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \{\mathrm{Pr}_i(j|\boldsymbol{\beta})\}^2 \boldsymbol{x}_{ijt_i}^T \boldsymbol{x}_{ijt_i} - \sum_{i=1}^{N} \sum_{t_i=1}^{T_i} \sum_{j=1}^{J} \mathrm{Pr}_i(j|\boldsymbol{\beta}) \boldsymbol{x}_{ijt_i}^T \boldsymbol{x}_{ijt_i},
\end{aligned}
$$

which is $R \times R$ Hessian matrix.

A.2 BHHH method

One of the alternative methods to Newton-Raphson method is BHHH method which would be introduced in this subsection. The Newton-Raphson method has two major drawbacks: calculating the Hessian is sometimes computation-intensive and it does not guarantee an increase in the log-likelihood if the log-likelihood is not globally concave.

The BHHH method uses matrix of outer products of the score as alternative to the negative Hessian in determining the next step. The score of an observation for consumer i, indexed by $s_i(\beta_r)$, is defined as the derivative of the observation's log-likelihood with respect to the parameter β_r which is in the form of

$$s_i(\beta_r) = \frac{\partial \ln(\Pr(j|\boldsymbol{\beta}))}{\partial \beta_r}.$$

Since the the log likelihood function for a standard logit model is written as

$$l(\boldsymbol{\beta}) \quad = \quad \sum_{j=1}^{J} y_{ij} \left\{ \boldsymbol{x}_{ijt_i}\boldsymbol{\beta} - \ln\left(\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) \right) \right\},$$

differentiating it with respect to vector β_r yields

$$s_i(\beta_r) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{j=1}^{J} y_{ij} \left\{ x_{ijt_ir} - \frac{\left(\frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r} \right)}{\sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\}, \tag{A.9}$$

since

$$\frac{\partial \boldsymbol{x}_{ijt_i}\boldsymbol{\beta}}{\partial \beta_r} = x_{ijt_ir}.$$

The last term in (A.9) is

$$\begin{aligned} \frac{\partial \sum_{l=1}^{J} \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r} &= \sum_{l=1}^{J} \frac{\partial \exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})}{\partial \beta_r} \\ &= \sum_{l=1}^{J} \left(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta}) x_{ilt_ir} \right). \end{aligned} \tag{A.10}$$

Substituting (A.10) back into (A.9) yields

$$\begin{aligned}
s_i(\beta_r) &= \sum_{j=1}^{J} y_{ij} \left\{ x_{ijt_ir} - \frac{\sum_{l=1}^{J}(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})x_{ilt_ir})}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \left\{ y_{ij}x_{ijt_ir} - \frac{y_{ij}\sum_{l=1}^{J}(\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})x_{ilt_ir})}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \left\{ y_{ij}x_{ijt_ir} - \frac{\exp(\boldsymbol{x}_{ijt_i}\boldsymbol{\beta})x_{ijt_ir}}{\sum_{l=1}^{J}\exp(\boldsymbol{x}_{ilt_i}\boldsymbol{\beta})} \right\} \\
&= \sum_{j=1}^{J} \{y_{ij}x_{ijt_ir} - \mathrm{Pr}_i(j|\boldsymbol{\beta})x_{ijt_ir}\} = \sum_{j=1}^{J} \{y_{ij} - \mathrm{Pr}_i(j|\boldsymbol{\beta})\}\, x_{ijt_ir}.
\end{aligned}$$

We repeat the above procedure for $r = 1, \cdots, R$ and stack them as $R \times 1$ vector which we denote $s_i(\boldsymbol{\beta})$. In the BHHH algorithm, the matrix $s_i(\boldsymbol{\beta})s_i(\boldsymbol{\beta})^T$ is used instead of negative of Hessian in Newton-Raphson method.

A.3 Variance of Estimates

The asymptotic covariance for correctly specified model is calculated as

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0, -\boldsymbol{H}^{-1})$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator, $\boldsymbol{\beta}^0$ denotes the true value of parameter and $\boldsymbol{H}$ is the expected Hessian in the population. The negative of this term $-\boldsymbol{H}$ is often called the information matrix (Train 2003). The asymptotic covariance of is $\hat{\boldsymbol{\beta}}$ is $-\boldsymbol{H}^{-1}/N$. In practice, the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is calculated as $-H^{-1}/N$ where H is the average Hessian in the sample. In calculating the asymptotic covariance of is $\hat{\boldsymbol{\beta}}$, W^{-1}/N and B^{-1}/N are used other than $-H^{-1}/N$, where W is the sample covariance of the scores and B is the sample average of outer product of the scores because it is known that $W \to -H$ as $N \to \infty$ and $B \to -H$ as $N \to \infty$ at the maximizing value of $\boldsymbol{\beta}$ by information identity (ibid).

For any model for which the expected score is zero at the true value is calculated as

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(0, \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1})$$

where $\boldsymbol{V}$ is the variance of scores in the population (ibid). The asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}/N$ in this case, and it is valued whether or not model is correctly specified or not. This matrix is called robust covariance matrix for this reason. In practice, $\boldsymbol{V}$ is substituted by W or B and the matrix is calculated as $H^{-1}WH^{-1}$. If model is correctly specified, $\boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}$ reduces to $-\boldsymbol{H}^{-1}$ since $-\boldsymbol{H}^{-1} = \boldsymbol{V}$ by information identity.

References

Bawa, Kapil (1990) 'Modeling inertia and variety seeking tendencies in brand choice behavior.' *Marketing Science* 9(3), 263–278

Berlyne, D. (1970) 'Novelty, complexity, and hedonic value.' *Journal of Marketing Research* 8(5), 279–286

Bucklin, Randolph E., Gary J. Russell, and V. Srinivasan (1998) 'A relationship between market share elasticities and brand switching probabilities.' *Journal of Marketing Research* 35(1), 99–113

Chintagunta, Pradeep K. (1998) 'Inertia and variety seeking in a model of brand-purchase timing.' *Marketing Science* 17(3), 253–270

Feinberg, Fred M., Barbara E. Kahn, and Leigh McAlister (1992) 'Market share response when consumers seek variety.' *Journal of Marketing Research* 29(2), 227–237

Givon, Moshe (1984) 'Variety seeking through brand switching.' *Marketing Science* 3(1), 1–22

Guadagni, Peter M., and John D. C. Little (1983) 'A logit model of brand choice calibrated on scanner data.' *Marketing Science* 2(3), 203–238

Gupta, Sachin, and Pradeep K. Chintagunta (1994) 'On using demographic variables to determine segment membership in logit mixture models.' *Journal of Marketing Research* 31(1), 128–136

Gupta, Sachin, Pradeep K. Chintagunta, and Dick R. Wittink (1997) 'Household heterogeneity and state dependence in a model of purchase strings: Empirical results and managerial implications.' *International Journal of Research in Marketing* 14(4), 341 – 357

Kamakura, Wagner A., and Gary J. Russell (1989) 'A probabilistic choice model for market segmentation and elasticity structure.' *Journal of Marketing Research* 26(4), 379–390

Keane, Michael P. (1997) 'Modeling heterogeneity and state dependence in consumer choice behavior.' *Journal of Business and Economic Statistics* 15(3), 310–327

Lattin, James M., and Leigh McAlister (1985) 'Using a variety-seeking model to identify substitute and complementary relationships among competing products.' *Journal of Marketing Research* 22(3), 330–339

McAlister, Leigh (1982) 'A dynamic attribute satiation model of variety-seeking behavior.' *Journal of Consumer Research* 9(2), 141–150

Seetharaman, P.B. (2003) 'Probabilistic versus random-utility models of state dependence: an empirical comparison.' *International Journal of Research in Marketing* 20(1), 87 – 96

— (2004) 'Modeling multiple sources of state dependence in random utility models: A distributed lag approach.' *Marketing Science* 23(2), 263–271

Train, K. (2003) *Discrete choice methods with simulation* (Cambridge Univ Pr)

Trivedi, Minakshi, Frank M. Bass, and Ram C. Rao (1994) 'A model of stochastic variety-seeking.' *Marketing Science* 13(3), 274–297