NUMERICAL OPTIMIZATION FOR MIXED LOGIT MODELS AND AN APPLICATION

A Thesis
Presented to
The Academic Faculty
by
Deniz Dogan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2008
NUMERICAL OPTIMIZATION FOR MIXED LOGIT MODELS AND AN APPLICATION

Approved by:

Dr. Anton Kleywegt, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Alexander Shapiro
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Shabbir Ahmed
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Sigrún Andradóttir
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Charles H. Rosa
General Motors Corporation
R & D and Planning

Date Approved: 16 Nov 2007
To my wife,

Mujde K. Dogan,

and my parents.
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my dissertation committee, Dr. Anton J. Kleywegt, Dr. Alexander Shapiro, Dr. Shabir Ahmed, Dr. Sigrún Andradóttir and Dr. Charles H. Rosa, for their time, valuable comments and suggestions.

In particular, I would like to thank my advisor, Dr. Anton Kleywegt, for his great guidance and support throughout my thesis. I would like to take this opportunity to express my gratitude to him for always having time for me whenever I had problems to discuss regarding both my thesis and my life. I really appreciated your advice and contributions that helped me go through those tough times.

Also, I feel extremely lucky to have met and worked with my co-advisor, Dr. Alexander Shapiro. His encouragement and guidance were undeniably effective in motivating me in those doubtful and hopeless times when progress was somewhat slow. So thank you for your continuous trust and support.

I would like to thank General Motors for supporting my research. Special thanks to Charles Rosa and Diane Gibbons, for being much more than work colleagues in those three internship summers in GM. Working with them taught me a lot, and in many aspects of life. Thanks for offering your help and friendship whenever I needed it.

My life in graduate school would not have been this memorable without my fellow friends in ISYE and Georgia Tech. Our lunch times together, board games played during weekends, soccer games, discussions about research problems were undoubtedly unforgettable. Thank you all for being great friends.

I also would like to thank my Mom, my Dad and my brother, for their love, support and especially for having faith in me, even at those times when I myself had
lost it.

Finally, I would like to thank my wife, Mujde. It would have been impossible to complete this dissertation without her. I want to thank you for standing by me and making me happy in my saddest moments just with your smile.
# TABLE OF CONTENTS

DEDICATION ................................................................. iii

ACKNOWLEDGEMENTS ...................................................... iv

LIST OF TABLES ........................................................... ix

LIST OF FIGURES .......................................................... x

SUMMARY ................................................................. xi

I  INTRODUCTION .......................................................... 1
   1.1 Overview of Discrete Choice Models .............................. 1
   1.2 Decision Maker ...................................................... 2
   1.3 Choice Set .......................................................... 2
   1.4 Utility Function ..................................................... 2
   1.5 Choice Probabilities ................................................. 3
   1.6 Types of Discrete Choice Models ................................ 3
      1.6.1 Multinomial Logit Models .................................. 3
      1.6.2 Nested Logit Models ......................................... 8
      1.6.3 Mixed Logit Models ......................................... 11
   1.7 Literature Review .................................................. 21

II ALGORITHM ............................................................. 25
   2.1 Motivation .......................................................... 25
   2.2 Trust-Region Algorithms .......................................... 28
      2.2.1 Trust-Region Algorithm Pseudocode ...................... 29
      2.2.2 Approximate Solution of the Subproblems ............... 29
   2.3 Errors ............................................................. 33
      2.3.1 Data Sample Error ........................................... 35
      2.3.2 Integration (Simulation) Sample Error .................... 36
      2.3.3 Error in Gradient Approximation ......................... 37
   2.4 Pseudocode ......................................................... 40
LIST OF TABLES

1.6.1 Mixed Logit - Substitution Pattern Demonstration ............... 19
2.7.1 CPU Times (sec) - AMLET .................................. 56
2.7.2 SLL Values for Simulated Problems ............................ 57
2.7.3 Relative Errors for Simulated Problems ....................... 57
2.7.4 SLL Values for Mobidrive Problems ........................... 59
2.7.5 Relative Errors for Mobidrive Problems ....................... 59
3.2.1 Low Segment - Market Share Data ............................. 66
3.4.1 Multinomial Logit Model - Results ............................ 77
3.4.2 Multinomial Logit Model - Market Share Results ............... 78
3.4.3 Multinomial Logit Model - Market Share Results Scenario 1 .... 79
3.4.4 Multinomial Logit Model - Penetration Rates for Cavalier .... 80
3.4.5 Mixed Logit Model # 1 - Results .............................. 83
3.4.6 Mixed Logit Model # 1 - Market Share Results ................ 84
3.4.7 Mixed Logit Model # 1 - Market Share Results Scenario ....... 88
3.4.8 Mixed Logit Model # 1 - Penetration Rates for Cavalier ....... 88
3.4.9 Mixed Logit Model # 2 - Results .............................. 89
3.4.10 Mixed Logit Model # 2 - Market Share Results ............... 91
3.4.11 Mixed Logit Model # 2 - Market Share Results Scenario ...... 93
3.4.12 Mixed Logit Model # 2 - Penetration Rates for Cavalier ....... 93
LIST OF FIGURES

2.7.1 Comparison of computation times in problem 2DIM. . . . . . . . . . . . 58
2.7.2 Comparison of computation times in problem 5DIM. . . . . . . . . . . . 58
2.7.3 Comparison of computation times in problem 10DIM. . . . . . . . . . . 58
2.7.4 Comparison of computation times in problem MDs. . . . . . . . . . . . 59
2.7.5 Comparison of computation times in problem MDc. . . . . . . . . . . . 60
SUMMARY

In this study, an algorithm (MLOPT) for mixed logit models is proposed. In comparison to other discrete choice models, mixed logit models are more flexible, but estimating mixed logit models with large datasets involves the solution of a nonlinear optimization problem with a high dimensional integral in the objective function, which is the log-likelihood function. This complex structure is a general problem that occurs in statistics and optimization. However, this study is particularly motivated by mixed logit models for demand modeling. Discrete choice models, specifically mixed logit models, have been successfully applied to demand modeling for many different industries.

Choice behavior is described in terms of a utility function. This function evaluates the utility each individual obtains from each alternative in his choice set by using the attributes of that alternative. A discrete choice model describes a setting in which each decision maker is faced with a finite choice set from which the alternative that maximizes the utility of the decision maker is selected. Let \( i \in \{1, \ldots, I\} \) be the index of the decision makers in the sample population and each decision maker \( i \) is faced with a choice set denoted by \( C_i \). Let \( j \) and \( l \) be indices for the alternatives in the choice sets \( C_i, i \in \{1, \ldots, I\} \). The actual alternative selected by the decision maker \( i \) is denoted by \( j_i \). Let \( U_{ij} \) be the utility that decision maker \( i \) derives from alternative \( j \). Then for \( i \in \{1, \ldots, I\} \), the following holds:

\[
U_{ij_i} \geq U_{ij}, \; \forall j \in C_i.
\]

For computational convenience, most of the time, the utility functions are assumed to be linear in the attributes. Let \( \mathbf{x}_{ij} \) be the vector of attributes related to the alternative \( j \) and the decision maker \( i \). The weights associated with these attributes
are represented by the vector, $\beta_i$. The vector $\beta_i$ is the relative importance given to each attribute by individual $i$. There is also a random part of the utility, which will be denoted by $\varepsilon_{ij}$. Hence, the utility function is modeled as:

$$U_{ij}(x_{ij}|\beta_i) := \beta_i'x_{ij} + \varepsilon_{ij}.$$  

The choice probability $L_{ij}$ is defined to be the probability that an individual $i$ chooses an alternative $j$ given the parameter vector $\beta_i$, i.e. it is the probability that the alternative under consideration has the largest utility in the given choice set.

$$L_{ij}(\beta_i) := \Pr(U_{ij} \geq U_{il}, \forall l \in C_i | \beta_i).$$

The density function for $\beta$, which is denoted by $f(\beta|\theta)$, needs to be estimated through the distribution parameters $\theta$, since the distribution of $\beta$ depends on $\theta$. The mixed logit choice probabilities, which will be denoted by $P_{ij}(\theta)$, are given as follows:

$$P_{ij}(\theta) := \int L_{ij}(\beta)f(\beta|\theta) \, d\beta.$$  

For each individual, the likelihood is defined as the choice probability of the alternative that is actually selected by the individual. Let $j_i$ denote the alternative selected by individual $i$. Hence, in mixed logit models, the objective function to be maximized is the summation of the log-likelihood functions for individuals in the dataset.

$$\max_{\theta} \left\{ LL(\theta) := \frac{1}{T} \sum_{i=1}^{I} \log \left( \int L_{ij_i}(\beta)f(\beta|\theta) \, d\beta \right) \right\}.$$  

In general, these choice probabilities are hard to evaluate due to the multidimensional integrals they contain. However, they can be approximated by Monte Carlo sampling. Using the Monte Carlo estimates of the choice probabilities, the simulated log-likelihood function can be constructed to estimate the log-likelihood function. Let $S_i$ be the integration sample for individual $i$, having size $|S_i|, \ i = 1, \ldots, I$. These are used to estimate the choice probabilities $P_{ij_i}(\theta)$ for each individual $i$. Given an integration sample $S = \{S_1, S_2, \ldots, S_I\}$, the choice probabilities $P_{ij}(\theta)$ can be estimated.
\[ SP_{ij}^S(\theta) := \frac{1}{|S_i|} \sum_{\nu \in S_i} L_{ij}(\beta^\nu(\theta)), \]

where \( SP_{ij}^S(\theta) \) is the simulated choice probability.

Finally, the simulated log-likelihood problem is denoted by:

\[
\max_\theta \left\{ SLL^S(\theta) := \frac{1}{T} \sum_{t=1}^T \log \left( SP_{ij}^S(\theta) \right) \right\}.
\]

The proposed algorithm, i.e. MLOPT, uses sampling from the dataset of individuals to generate a data sample. In addition to this, Monte Carlo samples are used to generate an integration sample to estimate the choice probabilities. Using the data sample and the integration sample, MLOPT estimates the log-likelihood function values for each individual in the dataset by controlling and adaptively changing the data sample and the size of the integration sample at each iteration. Another important feature of MLOPT is that it incorporates statistical testing for the quality of the solution obtained within the optimization problem.

Let \( N \) denote the data sample and \( S = \{ S_1, S_2, \ldots, S_T \} \) denote the integration sample. Using these samples, the simulated log-likelihood function can be rewritten as follows:

\[
\max_\theta \left\{ SLL^S_N(\theta) := \frac{1}{|N|} \sum_{i \in N} \log \left( SP_{ij}^S(\theta) \right) \right\}.
\]

In order to test MLOPT, a benchmark study from the literature that develops mixed logit demand models using both simulated datasets and a real-life dataset is used. These models are estimated using the MLOPT and the computational times are compared to those obtained by the mixed logit estimation tool from the literature (AMLET). The computational results are very encouraging.

As a next step, MLOPT is applied for real-life applications in the automotive industry. For this purpose, mixed logit demand models are constructed for predicting market shares in the Low Segment of the new car market. In order to demonstrate
the advantages of the mixed logit models, the same market segment is analyzed with a multinomial logit model using the same attributes.

The automotive industry is particularly interesting in that understanding the behavior of buyers and how rebates affect their preferences is very important for revenue management. Manufacturers spend billions of dollars on incentives to attract customers. However, without a good understanding of the effects of such incentives on market shares and eventually on profitability, these incentive programs may be misdirected.

Mixed logit demand models are particularly interesting because of their flexibility in constructing realistic substitution patterns. For instance, in a case where an additional rebate for an alternative (e.g., a rebate for an American car) is introduced, individuals having different car preferences (taste variations) will react differently to this rebate, meaning that an individual who likes American cars may be strongly affected by this change in scenario whereas an individual preferring European cars may not respond at all. Such behaviors that are incorporated in the model are called substitution patterns. In such realistic settings in which the respective behaviors of individuals under different scenarios are to be incorporated, there is an important advantage in using mixed logit models.

Real transaction data is used to generate and test the mixed logit models developed in this study. Another new aspect of this study is that the sales transactions are differentiated with respect to the transaction type of the purchases made. Hence, a vehicle model purchased by cash is treated as a different alternative than the same vehicle model purchased by financing or leasing. These mixed logit models are used to estimate demand and analyze market share changes under a different what-if scenario. These results comply with the estimates provided by the analysts in the industry.
CHAPTER I

INTRODUCTION

Discrete choice models are widely used for the modeling of individual choice behavior. They have played an important role in modeling travel demand. They performed very well in representing the travel behavior of individuals facing a finite number of alternatives. After this success, they have been successfully applied in various other areas such as predicting customer/buyer demand for the automobiles, grocery, personal computers, etc.

Among discrete choice models, multinomial logit models became very popular, especially because of their computational simplicity. However, they suffered from a very restrictive property, called independence from irrelevant alternatives (IIA), which will be described in later sections.

Models not having the above mentioned restriction exist, like probit and mixed logit models, but the computational effort required to solve these models made them undesirable. With the advent of simulation and the improvements in the CPU speeds, research on such models has increased. Among these, especially mixed logit models have been used extensively and have become popular.

1.1 Overview of Discrete Choice Models

Discrete choice models describe why or how a population of individuals make their choices among a finite number of alternatives. These models identify the importance that the individuals put on different attributes of the alternatives. They describe choices through a utility maximizing behavior of the individuals in the population and they model this behavior with given data.
1.2 Decision Maker

Decision makers are the individuals in the population of interest that are faced with choice decisions. In this study it will be assumed that the available data consist of at least a sample of the population, the alternative choices that they chose from, and the actual alternative that has been chosen by each individual.

1.3 Choice Set

The choice set is the set of alternatives from which a decision maker chooses one alternative. The main assumption of all the discrete choice models is that the cardinality of the choice set is finite.

1.4 Utility Function

Choice behavior is described in terms of a utility function. This function evaluates the utility each individual obtains from each alternative in his choice set by using the attributes of that alternative. These attributes may be specific to that alternative, like the fuel economy (miles per gallon) of the alternative, or they may be an interaction of the individual's attributes with the attributes of that alternative, e.g., the price of the alternative divided by the income of the decision maker.

Let \( i \in \{1, \ldots, I\} \) be the index of the decision makers in the set of observations and each decision maker \( i \) is faced with a choice set denoted by \( C_i \). The choice sets may be different for different decision makers. Let \( j \) and \( l \) be indices for the alternatives in the choice sets \( C_i, i \in \{1, \ldots, I\} \). The actual alternative selected by the decision maker \( i \) is denoted by \( j_i \). Let \( U_{ij} \) be the utility that decision maker \( i \) derives from alternative \( j \). Then for \( i \in \{1, \ldots, I\} \), the following holds:

\[
U_{ij_i} \geq U_{ij}, \quad \forall j \in C_i. \tag{1.4.1}
\]

For computational convenience, most of the time, the utility functions are assumed to be linear in the attributes. Let \( \mathbf{x}_{ij} \) be the vector of attributes related to the
alternative \( j \) and the decision maker \( i \). The weights associated with these attributes are represented by the vector, \( \beta_i \). The vector \( \beta_i \) is the relative importance given to each attribute by individual \( i \). There is also a random part of the utility having zero mean, which will be denoted by \( \varepsilon_{ij} \). Hence, the utility function is modeled as:

\[
U_{ij}(x_{ij} | \beta_i) := \beta_i' x_{ij} + \varepsilon_{ij}. \tag{1.4.2}
\]

### 1.5 Choice Probabilities

The decision maker \( i \) chooses an alternative \( j \) if the utility \( U_{ij} \) that is derived from that alternative is the highest among the alternatives in the choice set as shown in equation (1.4.1). The choice probability \( L_{ij} \) is defined to be the probability that an individual \( i \) chooses an alternative \( j \), i.e. it is the probability that the alternative under consideration has the largest utility in the given choice set.

Let \( L_{ij}(\beta_i) \) be the probability that the decision maker \( i \) chooses alternative \( j \) given the parameter vector \( \beta_i \).

\[
L_{ij}(\beta_i) := \Pr(U_{ij} \geq U_{il}, \forall l \in C_i | \beta_i).
\]

Equivalently,

\[
L_{ij}(\beta_i) = \Pr(\beta_i' x_{ij} + \varepsilon_{ij} \geq \beta_i' x_{il} + \varepsilon_{il}, \forall l \in C_i). \tag{1.5.1}
\]

### 1.6 Types of Discrete Choice Models

Depending on the assumptions on the distributions of the parameter vector \( \beta_i \), and the error term \( \varepsilon_{ij} \), the type of the discrete choice model is determined. Of the various discrete choice models that are available, the multinomial logit, nested logit, and mixed logit models are explained in detail for the purposes of this study.

#### 1.6.1 Multinomial Logit Models

The multinomial logit model has the least computational burden and hence is the most popular discrete choice model. One computational benefit comes from the fact
that the choice probabilities in equation (1.5.1) can be expressed in a simple form.

The main assumption of this model concerns the distributions of the error terms. In multinomial logit models, the error terms $\varepsilon_{ij}$ are assumed to be independent and identically (IID) Gumbel distributed. However, the IID property imposes a strong restriction on these models, and leads to the property of independence from irrelevant alternatives (IIA). The IIA property will be explained later in this section.

If the error term, $t$, is Gumbel distributed, then the cumulative distribution function is as follows:

$$F(t|\eta, \mu) := e^{-e^{-\mu(t-\eta)}} \quad , \quad \mu > 0,$$

and the probability density function is:

$$f(t|\eta, \mu) = \mu e^{-\mu(t-\eta)} e^{-e^{-\mu(t-\eta)}}.$$

In the above functions, the location parameter is $\eta$ and the scale parameter is $\mu > 0$.

Another restrictive assumption in multinomial logit models is the fact that the parameter vector, $\beta_i = \beta$, is assumed to be independent of $i$, meaning that the unknown parameter vector is constant over the population of decision makers. Thus equation (1.5.1) takes the following form:

$$L_{ij}(\beta) = \Pr(\beta' x_{ij} + \varepsilon_{ij} \geq \beta' x_{il} + \varepsilon_{il}, \forall l \in C_i).$$  \hspace{1cm} (1.6.1)

Under these assumptions, the choice probabilities $L_{ij}$ in equation (1.6.1) have a simple form (See Appendix A).

$$L_{ij}(\beta) = \frac{\exp(\beta' x_{ij})}{\sum_{l \in C_i} \exp(\beta' x_{il})}.  \hspace{1cm} (1.6.2)$$

The log-likelihood function is as follows:

$$LL(\beta) := \frac{1}{T} \sum_{i=1}^{T} \log(L_{iji}(\beta)).$$  \hspace{1cm} (1.6.3)$$

Then the maximum likelihood problem is given below:

$$\max_{\beta} \quad LL(\beta) = \frac{1}{T} \sum_{i=1}^{T} \log(L_{iji}(\beta)).$$  \hspace{1cm} (1.6.4)$$
Another desirable property of these models is that the log-likelihood problem shown in equation (1.6.4) is a concave unconstrained nonlinear maximization problem with continuously differentiable objective.

The most important restrictions of the Multinomial Logit models are summarized in the following subsections.

1.6.1.1 Proportional Substitutions

The proportional substitution property of the multinomial logit model can be explained as follows: Suppose that one of the alternatives in the choice set of every decision maker in the population becomes unavailable. In this case the market share of this ex-alternative will be captured by the remaining alternatives. The unfortunate outcome under multinomial logit models is that the market share of this ex-alternative will be shared proportionally among the remaining alternatives such that the ratios of the alternatives remain constant. This can be seen by considering the ratio of the choice probabilities of two alternatives \(j\) and \(l\):

\[
\frac{L_{ij}(\beta)}{L_{il}(\beta)} = \frac{\exp(\beta' x_{ij})}{\exp(\beta' x_{il})} = \exp(\beta' (x_{ij} - x_{il})).
\] (1.6.5)

The ratio of the choice probabilities does not depend on the attributes of the other alternatives. This is the restrictive independence from irrelevant alternatives (IIA) property of the multinomial logit model. This property can be further explained by using a term called elasticity. In general, elasticity is the percent change in one variable that is associated with a one percent change in another variable. In discrete choice models, the elasticity of choice probability (market share) is considered with respect to changes in one of the attributes of an alternative. For instance, in a scenario in which there is a change in an attribute of an alternative, such as a decrease in its price, it is expected that the market share or choice probabilities of the alternative under consideration will increase, i.e., it will capture market shares from the other alternatives. The new market shares under this new scenario will depend on the
elasticities of the other alternatives with respect to the change in the price of the considered alternative.

Let \( i \) be an index for the set of individuals \( \{1, 2, \ldots, I\} \), and let \( j, k, l \) be the indices for different alternatives. For individual \( i \), the elasticity of the choice probability of alternative \( j \) associated with one percent change in the \( n^{th} \) attribute of alternative \( k \) will be denoted by \( E_{ijx^n_k} \).

The elasticity of market share, \( E_{ijx^n_k} \), for \( j = k \) is calculated as follows:

\[
E_{ijx^n_j} := \frac{\partial}{\partial x^n_{ij}} L_{ij}(\beta) \times \frac{x^n_{ij}}{L_{ij}(\beta)} \\
= \beta^n \left( \frac{\exp(\beta' x_{ij})}{\left( \sum_{l \in C_i} \exp(\beta' x_{il}) \right)^2} - \left( \frac{\exp(\beta' x_{ij})}{\sum_{l \in C_i} \exp(\beta' x_{il})} \right)^2 \right) \times \frac{x^n_{ij}}{L_{ij}(\beta)} \\
= \beta^n x^n_{ij} (1 - L_{ij}(\beta)).
\]

The elasticity of market share, \( E_{ijx^n_{ik}} \) for \( j \neq k \), also called cross-elasticity, is calculated as follows:

\[
E_{ijx^n_{ik}} := \frac{\partial}{\partial x^n_{ik}} L_{ij}(\beta) \times \frac{x^n_{ik}}{L_{ij}(\beta)} \\
= \left( \frac{\exp(\beta' x_{ij})}{\left( \sum_{l \in C_i} \exp(\beta' x_{il}) \right)^2} \exp(\beta' x_{ik}) \beta^n \right) \times \frac{x^n_{ik}}{L_{ij}(\beta)} \\
= -\beta^n x^n_{ik} L_{ik}(\beta). \quad (1.6.6)
\]

When the \( n^{th} \) attribute of alternative \( k \) is increased by one percent, all the other alternatives \( j \neq k \) will be affected by the same ratio since their cross-elasticities (1.6.6) will only depend on alternative \( k \). This is the restatement of the restrictive independence from irrelevant alternatives (IIA) property of the multinomial logit model. Since the cross-elasticities are equal to each other, their market shares decrease proportionally.
1.6.1.2 Taste Variations

Taste variations are an important part of demand models. Most of the time, the decision makers in the population will have different tastes for the attributes of the alternatives or the alternatives as a whole. There are two types of taste variations: \textit{systematic taste variations} and \textit{random taste variations}.

Systematic taste variations are variations that can be explained by a function of some attributes of the decision makers themselves, such as education level or income level. These are incorporated in multinomial logit models by using attributes that depend on the attributes of the individuals. For example, the $\mathbf{\beta}$ vector can be a function of some attributes of the decision maker. Multinomial logit models cannot incorporate this type of taste variation directly, because in these models the parameter vector $\mathbf{\beta}$ is the same for every individual. Instead, an alternative approach is suggested in the literature to incorporate this type of taste variation in multinomial logit models, which makes use of the attributes that depend on the decision makers. For example, let $Income_i$ be the income level of decision maker $i$, $p_j$ be the price of the alternative $j$. In order to model the systematic taste variation in multinomial logit models, a new variable having a constant coefficient should be defined. Letting this constant coefficient be $\alpha$, suppose the relative importance that individuals give to price is inversely proportional to their income levels. Then the new variable can be denoted as $p_j/Income_i$. Note that this new variable can be used in the utility function in multinomial logit models since its coefficient $\alpha$ will be unknown but constant over the decision makers. Nevertheless, the real coefficient for price will depend on the income level of the decision maker $i$.

$$\alpha(p_j/Income_i) = (\alpha/Income_i)p_j = \alpha_4 p_j,$$

where $\alpha_i = \alpha/Income_i$ is the actual coefficient for price and it depends on the income level of the decision makers.
The second type of taste variation is the random taste variation that cannot be modeled systematically. Such cases occur when, for instance, different individuals with the same attributes have different tastes for the attributes of the alternatives. One such attribute could be, for instance, price. In these cases, the best way to explain these variations is to use random coefficients $\beta$ in the utility function model for those attributes. However, this is not possible in multinomial logit models. Mixed logit models, which will be explained in Section 1.6.3, are not only able to handle this latter issue, but are also able to incorporate individual systematic taste variations for each decision maker.

1.6.2 Nested Logit Models

Nested logit models appear in the literature as an intermediary step between multinomial logit models and the mixed logit models (see Chapter 10 of Ben-Akiva and Lerman (1985) [10]). In nested logit models the set of alternatives are partitioned into a set of groups called nests. These nests are constructed a priori by the researcher. In a nested logit model, the alternatives in each nest are correlated with each other, but are not correlated with the alternatives in the other nests. Accordingly, when a scenario is applied, such as an additional rebate for an alternative, the other alternatives that are also present in that alternative’s nest are affected more than the alternatives that are not in the same nest.

The notation and example depicted below is adopted from Ben-Akiva and Lerman (1985) [10]. Suppose the finite set of alternatives $C$ can be composed by the use of two attributes, mode ($m$) and destination ($d$). Hence, each different combination of ($m, d$) pair denotes an alternative. The observed part of the utility that is shared by each mode $m$ is denoted by $V_m$. Likewise, the observed part of the utility that is shared by each destination $d$ is denoted by $V_d$. Let $V_{md}$ be part of the observed utility that is specific to each alternative. In nested logit models, the unobserved part of the
utility is also assumed to have such categorization \((\varepsilon_m, \varepsilon_d, \text{and } \varepsilon_{md})\). Under these assumptions, the utility function \(U_{md}\) becomes:

\[
U_{md} = V_m + V_d + V_{md} + \varepsilon_m + \varepsilon_d + \varepsilon_{md},
\]

where

\[
V_m = \beta x_m, \ V_d = \alpha y_d, \ V_{md} = \gamma \varepsilon_{md}.
\]

In nested logit models, either \(\varepsilon_m\) or \(\varepsilon_d\) has zero variance. In this example, suppose \(\text{var}(\varepsilon_d) = 0\). In addition, there are also the following assumptions on the distribution of error terms:

- The error terms \(\varepsilon_m\) and \(\varepsilon_{md}\) are independent \(\forall m, d\).
- The error terms \(\varepsilon_{md}\) are IID Gumbel distributed with scale parameter \(\mu^d\).
- The error terms \(\varepsilon_m\) are distributed such that \(\max_{vd} U_{md}\) is Gumbel distributed with scale parameter \(\mu^m\).

Using these assumptions, the constructed nested logit model assumes a two-level decision tree structure, where in the first level, the decision maker chooses mode \((m)\), after which the destination \((d)\) is chosen. In this tree, each different mode \((m)\) is considered to form a nest. The marginal choice probability for each nest \((\text{mode})\) is calculated as follows:

\[
P(m) = \frac{\exp \left((V_m + \tilde{V}_m) \mu^m \right)}{\sum_{m'} \exp \left((V_{m'} + \tilde{V}_{m'}) \mu^{m'} \right)},
\]

where

\[
\tilde{V}_m = \frac{1}{\mu^d \log} \left(\sum_{d} \exp (V_d + V_{dm}) \right).
\]

In order to calculate the choice probability of a specific alternative \((m, d)\), the conditional probability of choosing destination \(d\), given that mode \(m\) is chosen, needs
to be calculated as well:

\[
P(d|m) = \frac{\exp \left( (V_{md} + V_d) \mu^d \right)}{\sum_{d'} \exp \left( (V_{md'} + V_{d'}) \mu^{d'} \right)}.
\]

Hence, the choice probability of alternative \((m, d)\) can be written as:

\[
P(m, d) = P(m) \times P(d|m).
\]

Note that only the ratio \(\mu^m / \mu^d\) can be estimated. See Ben-Akiva and Lerman (1985) [10] for the derivation of the probabilities.

The properties of nested logit models related to substitution patterns are summarized in Train (2002) [40] as:

- For any two alternatives that are in the same nest, the ratio of probabilities is independent of the attributes or existence of all other alternatives. Hence, the IIA property holds among the alternatives within the same nest, as is the case in multinomial logit models.

- For any two alternatives in different nests, the ratio of probabilities can depend on the attributes of the other alternatives in the two nests. This is the generalization over the multinomial case, since the IIA property does not hold between the alternatives, belonging to different nests.

These claims can be justified by analyzing the ratio of choice probabilities between any two alternatives that belong to the same nest (same mode) and the ratio of choice probabilities between any two alternatives that are from separate nests (different modes).

Let \(md\) and \(md'\) be two different alternatives that belong to the same nest, namely \(m\). Then the ratio of the choice probabilities are as follows:

\[
\frac{P(m, d)}{P(m, d')} = \frac{\exp (V_{md})}{\exp (V_{md'})}.
\]
Note that the above ratio is the same ratio as for a multinomial logit model. Furthermore, this ratio only depends on the attributes of those two alternatives. Hence, this proves that within the nests the substitution pattern suffers from the IIA property.

Let \( md \) and \( m'd' \) be two alternatives that are from two different nests \((m \text{ and } m')\). Then the ratio of the choice probabilities are as follows:

\[
\frac{P(m, d)}{P(m', d')} = \frac{\exp \left((V_m + V_m')\mu^m \right) \exp \left((V_{md} + V_{d'})\mu^d \right)}{\exp \left((V_{m'} + V_{m'})\mu^{m'} \right) \exp \left((V_{m'd'} + V_{d'})\mu^{d'} \right)}.
\]

Note that this ratio depends not only on the attributes of these two alternatives, but also on the attributes of the other alternatives in their respective nests. However, this ratio is independent of the attributes of the nests other than those two nests containing \(d\) and \(d'\).

The most immediate drawback of the nested logit models is that the researcher has to construct these nests beforehand. However, it is hard to create nests with respect to continuous attributes such as Operating Cost or NPV. Another problem related to nested logit models is that the effect of the additional rebate will be at the same level for the alternatives within the same nest, just like in the multinomial logit model. Therefore, on the inter-nest level (i.e. between different nests), nested logit models are similar to mixed logit models in that they do not exhibit the IIA property. However, on the intra-nest level (i.e. within a particular nest), they behave like multinomial logit models.

1.6.3 Mixed Logit Models

To overcome some shortcomings of the multinomial logit model, various generalizations were introduced in the literature. For example, to overcome the IIA behavior described previously, the nested logit model was introduced (See Section 1.6.2). Nested logit models have simple expressions for the choice probabilities but still cannot represent the many forms of random taste variations. The generalization that will be the core of this study is the mixed logit model, also called the logit kernel model, which is
a further generalization of the nested logit model. These models have the advantage of greater generality, and thus greater potential realism, but at the expense of being much less tractable than the multinomial logit model. The mixed logit model is seen as the model of the future (Walker, Ben-Akiva, and Bolduc (2007) [42]), and has become very popular with the development of simulation methods. Before discussing the computational and analytical challenges introduced by the mixed logit model, a brief motivation and description of these models is given first.

Mixed logit models are powerful models, which can approximate any utility maximizing discrete choice model as closely as possible under mild regularity conditions (McFadden and Train (2000) [27]). For example, Brownstone and Train (1999) [18] and Ben-Akiva and Bolduc (1996) [9] approximated probit model using mixed logit model and report that the approximations are as accurate as direct simulation methods for probit model. Brownstone and Train (1999) [18] also describe approximating a nested logit model using a mixed logit model. The advantages of mixed logit models over multinomial logit models are their ability to incorporate random taste variations, unrestricted substitution patterns, and correlations in unobserved factors over time (Train (2002) [40]).

Of course this flexibility comes at a price. Estimating mixed logit models with large datasets involves the solution of a nonlinear optimization problem with a high dimensional integral in the objective function. This type of structure occurs not only in mixed logit estimation problems, but also in other problems in statistics and optimization such as Stochastic Optimization.

As already mentioned in Section 1.6.1.2, the multinomial logit models cannot incorporate random taste variations and they can only account for systematic taste variations with a constant parameter vector $\beta$ for all individuals. In mixed logit models, however, it is assumed that the parameter vector $\beta$ is not constant, but rather is random over the population of decision makers and has a known type of probability
distribution with parameter vector $\theta$ to be estimated. For instance, for multivariate normal distributions, $\theta$ denotes the mean vector and the covariance matrix of $\beta$.

As regards the proportional substitution property, mixed logit models do not exhibit such a restrictive behavior. This issue will be discussed in the next subsection.

Let $\tilde{\beta}$ be a realization of $\beta$. Given $\tilde{\beta}$, the conditional choice probability is the same as in multinomial logit models. Then the conditional choice probability, $L_{ij}(\tilde{\beta})$, is as given in equation (1.6.2).

In order to calculate the mixed logit choice probabilities, the density function for $\beta$, which is denoted by $f(\beta|\theta)$, needs to be estimated through the distribution parameters $\theta$, since the distribution of $\beta$ depends on $\theta$. The mixed logit choice probabilities, which will be denoted by $P_{ij}(\theta)$, are given as follows:

$$P_{ij}(\theta) := \int L_{ij}(\beta)f(\beta|\theta) \, d\beta.$$  \hspace{1cm} (1.6.7)

Using the choice probabilities, the log-likelihood function is defined as follows:

$$LL(\theta) := \frac{1}{I} \sum_{i=1}^{I} \log \left( \int L_{ij}(\beta)f(\beta|\theta) \, d\beta \right).$$  \hspace{1cm} (1.6.8)

Here, instead of maximizing the log-likelihood function over $\beta$, it should be maximized over the distribution parameters, $\theta$:

$$\max_{\theta} \left\{ LL(\theta) = \frac{1}{I} \sum_{i=1}^{I} \log \left( \int L_{ij}(\beta)f(\beta|\theta) \, d\beta \right) \right\}.$$  \hspace{1cm} (1.6.9)

Each choice probability in the log-likelihood function (1.6.8) involves a multidimensional integral. Typically, these integrals cannot be calculated exactly, but can be approximated with Monte Carlo estimates. For this purpose, random samples of $\beta$ are generated for each individual. These random samples are called integration (simulation) samples, $S_i$, for each individual $i$, and they are represented by $\{\beta^\nu, \nu \in S_i\}$. Suppose the relative importance vector $\beta \in \mathbb{R}^K$ is distributed normal with the means vector $\mu$ and the covariance matrix $\Sigma$:

$$\beta \sim N(\mu, \Sigma).$$

13
The covariance matrix can be rewritten as:

$$
\Sigma = Q'Q,
$$

where $Q'Q$ is the upper triangular Cholesky factorization of $\Sigma$. In this case, the parameter vector is given by:

$$
\theta = \{ \mu_1, \mu_2, \ldots, \mu_K, q_{11}, q_{12}, \ldots, q_{1K}, q_{22}, q_{23}, \ldots, q_{2K}, \ldots, q_{KK} \},
$$

where $q_{ij}$ is an element of the upper triangular matrix $Q$.

In order to generate the random sample $\{ \beta^\nu \}$, an independent random sample of standard normals is generated for each component of $\beta$. This sample, which is denoted by $\{ \xi_1^\nu, \ldots, \xi_K^\nu \}$, is used to construct the components of the vector $\beta^\nu(\theta)$:

$$
\beta^\nu_k(\theta) = \mu_k + \sum_{j=k}^{K} \xi_j^\nu q_{kj}, \forall k = 1, \ldots, K, \forall \nu \in S_i.
$$

Hence, $S_i$ is the integration sample for individual $i$, having size $|S_i|$, $i = 1, \ldots, I$. These are used to estimate the choice probabilities $P_{ij}(\theta)$ for each individual $i$. Given an integration sample $S = \{ S_1, S_2, \ldots, S_I \}$, the choice probabilities $P_{ij}(\theta)$ in (1.6.7) are estimated by:

$$
SP_{ij}^{S_i} := \frac{1}{|S_i|} \sum_{\nu \in S_i} L_{ij}(\beta^\nu(\theta)),
$$

where $SP_{ij}^{S_i} (\theta)$ is the simulated choice probability.

The log-likelihood function $LL(\theta)$ in (1.6.8) can be estimated by:

$$
SLL^S(\theta) := \frac{1}{I} \sum_{i=1}^{I} \log \left( SP_{ij}^{S_i}(\theta) \right),
$$

where $SLL^S(\theta)$ is the simulated log-likelihood function.

Finally, the problem (1.6.9) is approximated by:

$$
\max_{\theta} \left\{ SLL^S(\theta) = \frac{1}{I} \sum_{i=1}^{I} \log \left( SP_{ij}^{S_i}(\theta) \right) \right\}.
$$
1.6.3.1 Substitution Pattern in Mixed Logit Models

Multinomial logit models have the restrictive proportional substitution pattern. As explained in the previous section, one of the reasons for this property is the fact that the cross-elasticities of the choice probabilities are equal. Mixed logit models do not have the IIA property, since the elasticities of the choice probabilities for different alternatives need not be the same. In this section, the elasticities in mixed logit models will be discussed and their effects on the substitution pattern will be explained.

Suppose the relative importance vector $\mathbf{\beta}$ is random having a density function $f(\mathbf{\beta}|\mathbf{\theta})$, where $\mathbf{\theta}$ is the parameter vector of the distribution. This is also called mixing density function in the literature. For individual $i$, the choice probability of alternative $j$ is denoted by $P_{ij}(\mathbf{\theta})$ in (1.6.7).

The cross-elasticity of choice probability for alternative $j$ ($j \neq k$) when the $n^{th}$ attribute of alternative $k$ is increased by one percent is denoted by $E_{ijx_{ik}^n}$ and is calculated as follows:

$$E_{ijx_{ik}^n} := \frac{\partial}{\partial x_{ik}^n} P_{ij}(\mathbf{\theta}) \times \frac{x_{ik}^n}{P_{ij}(\mathbf{\theta})}$$

$$= \left( \int - \frac{\exp(\mathbf{\beta}'x_{ij})}{\left( \sum_{l \in C_i} \exp(\mathbf{\beta}'x_{il}) \right)^2} \exp(\mathbf{\beta}'x_{ik}) \mathbf{\beta}^n f(\mathbf{\beta}) d\mathbf{\beta} \right) \times \frac{x_{ik}^n}{P_{ij}(\mathbf{\theta})}$$

$$= -x_{ik}^n P_{ik}(\mathbf{\theta}) \int \mathbf{\beta}^n \frac{L_{ik}(\mathbf{\beta})}{P_{ik}(\mathbf{\theta})} \frac{L_{ij}(\mathbf{\beta})}{P_{ij}(\mathbf{\theta})} f(\mathbf{\beta}) d\mathbf{\beta}. \quad (1.6.13)$$

Likewise, the cross-elasticity for alternative $l$ ($l \neq k$), denoted by $E_{ilx_{ik}^n}$, is calculated as follows:

$$E_{ilx_{ik}^n} := -x_{ik}^n P_{ik}(\mathbf{\theta}) \int \mathbf{\beta}^n \frac{L_{ik}(\mathbf{\beta})}{P_{ik}(\mathbf{\theta})} \frac{L_{il}(\mathbf{\beta})}{P_{il}(\mathbf{\theta})} f(\mathbf{\beta}) d\mathbf{\beta}. \quad (1.6.14)$$

Consider a situation where it is believed that an improvement in alternative $k$ draws relatively more market share from alternative $j$ than from alternative $l$. 

15
Train (2002) [40] suggests specifying an element of the attribute vector \( \mathbf{z} \) such that this element is similar for alternatives \( k \) and \( j \) but not for alternatives \( k \) and \( l \), where the attribute has a random coefficient. Therefore, the correlation of the conditional choice probabilities \( L_{ik}(\beta') \) and \( L_{ij}(\beta') \) is higher than the correlation of the conditional choice probabilities \( L_{ik}(\beta') \) and \( L_{il}(\beta') \). Due to this fact, the integral in the cross-elasticity equation (1.6.13) for alternative \( j \) will be higher in magnitude than the integral in the cross-elasticity equation (1.6.14) for alternative \( l \). Accordingly, cross-elasticity is higher in magnitude for alternative \( j \) than for alternative \( l \), resulting in the fact that alternative \( j \) is more heavily affected than alternative \( l \).

The concepts discussed above can be demonstrated by the following example. Suppose that the substitution pattern that is believed to exist between the alternatives depends only on the operating cost attribute. In other words, an improvement for one of the alternatives will affect the alternatives having similar operating costs more than the alternatives having different operating costs. In order to construct this substitution pattern in mixed logit models, the coefficient of the operating cost attribute will be assumed to be random.

Suppose that alternatives \( k \) and \( j \) have similar operating costs whereas alternative \( l \) has a different operating cost. In this scenario, the price of alternative \( k \) is changed. For the simplicity of calculations the coefficient of the operating cost attribute is the only random coefficient. The cross-elasticity equation (1.6.13) for alternative \( j \) becomes:

\[
E_{ij_{price}} = -\beta_{price} x_{ij_{price}} P_{ik}(\theta) \left( 1 + \text{Cov} \left( \frac{L_{ik}(\beta)}{P_{ik}(\theta)}, \frac{L_{ij}(\beta)}{P_{ij}(\theta)} \right) \right).
\]

Likewise, the cross-elasticity equation (1.6.14) for alternative \( l \) becomes:

\[
E_{il_{price}} = -\beta_{price} x_{il_{price}} P_{ik}(\theta) \left( 1 + \text{Cov} \left( \frac{L_{ik}(\beta)}{P_{ik}(\theta)}, \frac{L_{il}(\beta)}{P_{il}(\theta)} \right) \right).
\]
Note that the cross-elasticity formula for $E_{ijx_{ik}}^{\text{price}}$ is the multinomial logit cross-elasticity multiplied by $\text{Cov} \left( \frac{L_{ik}(\beta)}{P_{ik}(\theta)}, \frac{L_{ij}(\beta)}{P_{ij}(\theta)} \right) + 1$. In the extreme case, when the parameters are constant, then the covariance term is zero and the aforementioned cross-elasticity becomes the multinomial logit model’s cross-elasticity. On the other hand, if the covariance term between alternative $j$ and $k$ is high, then the cross-elasticity between them is higher than the cross-elasticity between $l$ and $k$.

Since alternatives $j$ and $k$ have similar operating costs and alternative $l$ has a different operating cost,

$$\text{Cov} \left( \frac{L_{ik}(\beta)}{P_{ik}(\theta)}, \frac{L_{ij}(\beta)}{P_{ij}(\theta)} \right) > \text{Cov} \left( \frac{L_{ik}(\beta)}{P_{ik}(\theta)}, \frac{L_{il}(\beta)}{P_{il}(\theta)} \right).$$

The reason for this is as follows. Both alternative $j$ and alternative $k$ have higher Operating Cost attribute values. Hence, for smaller $\beta^{OC}$ values, the likelihood function values $(L_{ij}(\beta), L_{ik}(\beta))$ will be relatively lower, whereas $L_{il}(\beta)$ will be higher. Similarly, the likelihood values $(L_{ij}(\beta), L_{ik}(\beta))$ will be relatively higher for higher $\beta^{OC}$ values compared to $L_{il}(\beta)$.

Therefore $|E_{ijx_{ik}}^{\text{price}}| > |E_{ilx_{ik}}^{\text{price}}|$. Since the magnitude of the cross-elasticity of alternative $j$ is higher than that of alternative $l$, alternative $j$ is more sensitive to the price change of alternative $k$.

Furthermore, as the standard deviations of the random terms increase, the aforementioned covariances and thus the elasticities also increase. In the extreme case, when the standard deviation is zero, the covariance terms in the cross-elasticity equations are also equal to zero, i.e. $\beta$ is constant, and the cross-elasticities are equal for all the alternatives. Here, the same situation as in the multinomial logit case occurs and the proportional substitution pattern is observed:

$$E_{ijx_{ik}}^{\text{price}} = E_{ilx_{ik}}^{\text{price}} = -\beta^{\text{price}} x_{ik}^{\text{price}} P_{ik}(\theta).$$

Consider a small numeric example given in Sandor (2001) [36]. Suppose there are
two alternatives and three attributes. Their attribute vectors are given below:

\[
x_1 = (1, 1, 1) \\
x_2 = (0.5, 1.5, 1).
\]

Assume that a multinomial logit model is estimated and the relative importance vector \( \beta \) is given by:

\[
\beta = (1, 1, 1).
\]

Then, the market shares of these alternatives \( s_1 \) and \( s_2 \) are easily calculated:

\[
s_1 = s_2 = 0.5.
\]

Under a new scenario, assume that a new alternative with attribute vector:

\[
x_3 = (1, 1, 1)
\]

is introduced. Then the new market shares under this scenario become:

\[
s_1 = s_2 = s_3 = 1/3.
\]

Note that because of the unrealistic substitution pattern of the multinomial logit model, both of the original alternatives lose equal percent market shares under this scenario.

In order to compare this situation with mixed logit models, suppose the components of \( \beta \) are independently normally distributed with mean vector \( \mu = (1, 1, 1) \) and the standard deviations \( (\sigma, \sigma, \sigma) \).

For the base case, when there are two alternatives as described before, the market shares do not depend on \( \sigma \):

\[
s_1 = s_2 = 0.5 \forall \sigma \geq 0
\]

However, when the third alternative is introduced, the market shares will depend on the standard deviations of the parameters. Table 1.6.1 gives the resulting market
Table 1.6.1: Mixed Logit - Substitution Pattern Demonstration

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32.5</td>
<td>35</td>
<td>32.5</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>26.5</td>
<td>47</td>
<td>26.5</td>
</tr>
<tr>
<td>30</td>
<td>25.5</td>
<td>49</td>
<td>25.5</td>
</tr>
</tbody>
</table>

...shares for the three alternatives with different \( \sigma \) values. By increasing the standard deviations of the parameters, the effect of introducing the new alternative over the second alternative becomes negligible as expected. This is due to the fact that the attribute values for the first and the third alternatives are the same, hence these alternatives are perfect substitutes. On the other hand if the standard deviation of the parameters are close to zero, then the model behaves just like the multinomial logit model. Therefore, introducing a new alternative captures equal percent of market share from the existing alternatives.

In order to give a more complex example, suppose it is believed that the consumers have preferences for vehicles that they purchase based on the vehicle’s region of origin (e.g. American, Asian, European). This implies a substitution pattern, in which the cross-elasticities between vehicles that belong to the same region of origin are higher than the cross-elasticities between vehicles that do not belong to the same region of origin.

Using mixed logit models, this substitution pattern can be constructed as follows. It will be assumed that the attributes of the utility function consist of vehicle model specific indicators with random coefficients and other attributes with constant coefficients. Let \( C = \{1, 2, \ldots, |C|\} \) be the set of all vehicle models and let US, AS, and EU be non-overlapping subsets (based on region of origin) of \( C \) such that \( \text{US} \cup \text{AS} \cup \text{EU} = C \). Assume that the parameters for the vehicle model specific indicators are represented by \( \gamma \in \mathbb{R}^{|C|} \). This vector can be decomposed into three
parts depending on the region of origin:

\[
\gamma = \begin{pmatrix}
\gamma^{\text{US}} \\
\gamma^{\text{AS}} \\
\gamma^{\text{EU}}
\end{pmatrix}.
\]

To create the required substitution behavior, the distributions of \(\gamma^{\text{US}}, \gamma^{\text{AS}},\) and \(\gamma^{\text{EU}}\) will be independent and are given below:

\[
\begin{aligned}
\gamma^{\text{US}} &\sim N(\mu^{\text{US}}, \Omega^{\text{US}}), \\
\gamma^{\text{AS}} &\sim N(\mu^{\text{AS}}, \Omega^{\text{AS}}), \\
\gamma^{\text{EU}} &\sim N(\mu^{\text{EU}}, \Omega^{\text{EU}}),
\end{aligned}
\]

where \(\mu^{\text{US}}, \mu^{\text{AS}},\) and \(\mu^{\text{EU}}\) are the mean vectors and \(\Omega^{\text{US}}, \Omega^{\text{AS}},\) and \(\Omega^{\text{EU}}\) are the full covariance matrices for the parameters \(\gamma^{\text{US}}, \gamma^{\text{AS}},\) and \(\gamma^{\text{EU}},\) respectively.

If it is expected that a change in any one of the attributes (such as a reduction in price) of an American vehicle model affects the other American vehicle models more than the European and the Asian vehicle models, then the off-diagonal elements of \(\Omega^{\text{US}}, \Omega^{\text{AS}},\) and \(\Omega^{\text{EU}}\) need to be positive. This means that the cross-elasticities between the American vehicle models are higher in magnitude than the cross-elasticities between the American vehicle models and the vehicle models from other regions of origin. As explained in the example with a random coefficient for the Operating Cost attribute, the magnitudes of the cross-elasticities are proportional to the covariances between the likelihoods of the different alternatives. In this example, \(\gamma^{\text{US}} \sim N(\mu^{\text{US}}, \Omega^{\text{US}})\) and the positive off-diagonal elements of the covariance matrix \(\Omega^{\text{US}}\) imply that there is a positive correlation between the components of the vector \(\gamma^{\text{US}}.\) This correlation increases the covariances of the likelihoods and naturally the cross-elasticities between the American vehicle models. Similar arguments hold for European and Asian vehicle models.
1.7 Literature Review


The availability and the improvements in computer power triggered the use of mixed logit models on customer-level data. These models were capable of including customer heterogeneity. Brownstone and Train (1999) [18] estimated a mixed logit model from a stated-preference survey for customers’ preferences among vehicles having different fuel types. A study by Bhat and Castellar (2002) [14] estimated a mixed logit model for transportation mode and time of day choice for using the San Francisco Bay Bridge. They used quasi-Monte Carlo draws to simulate the log-likelihood function. Revelt and Train (1998) [34] estimated the impact of rebates and loans on residential customers’ choice of efficiency level for refrigerators. In their study, they accounted for the correlations in the unobserved utility using random parameters. Hensher, Shore, and Train (2005) [26] evaluated how much customers are willing to pay for specific levels of drinking water service. They used stated preference data to estimate mixed logit models.

Mixed logit models have been used in different markets and industries. Revelt and Train (1998) [34], McFadden and Train (2000) [27] used them for the energy

Hensher and Greene (2003) [25] reviewed the progress in the estimation of mixed logit models. They emphasized the properties of the mixed logit model that differentiate them from other models and presented some important issues that should be taken into account in mixed logit model estimation, such as selection of the random parameters and their distributions, preference heterogeneity, correlated choice situations and correlations between parameters. They stated that mixed logit models require high quality data as input due to their ability to incorporate customer heterogeneity in choice decisions.

Chiou and Walker (2005) [21] discussed one of the important problems observed while estimating mixed logit models, namely the identification problem. They presented a mixed logit model from a real-life dataset of households’ choices over telephone services. They first showed that the mixed logit model built for the dataset was theoretically unidentifiable, but on the other hand, they observed that with a small integration sample size the model seemed identifiable. Only a large sample size revealed the unidentifiability implicit in the model. They also investigated a theoretically identifiable but empirically unidentifiable mixed logit model for DVD store purchases. They reported that the identification problem may be hidden when small integration sample sizes are used. They observed that this problem reveals itself with a sufficient integration sample size by either exploding parameter estimates or singular Hessian matrices.

McFadden and Train (1995) [28] examined the integration sample variance in
the estimated parameters for mixed logit models using Monte Carlo sampling. The analysis was based on data regarding households’ choices among classes of new cars to buy.

Train and Winston (2007) [41] analyzed the vehicle choice behavior in the U.S. automotive market. They developed a mixed logit model and handled price endogeneity using the BLP approach developed in a study by Berry, Levinsohn, and Pakes (1995) [12]. They considered each vehicle model as an alternative, and hence did not consider the type of purchase or the length of the purchase contract in their analysis. Dasgupta, Siddarth, and Silva-Risso (2007) [22] also analyzed the vehicle choice behavior. They developed a nested logit model for predicting the demand for vehicle models with leasing and financing options. The data consisted of individual level transactions of new cars from the lower luxury segment in the Southern California region. They considered the contract term lengths for lease and finance separately and an alternative consisted of the vehicle model, transaction type (finance or lease) and the contract term (e.g., 24, 36, 48 and 60-month terms for financing contracts; 36 and 48-month terms for lease contracts). The price of the alternative was calculated using the present value of the investment. A more detailed description of the development and implementation of the PIN Incentive Planning System was given in a study by Silva-Risso and Ionova (2007) [37]. The system is currently being used by most major automobile manufacturers. It has been credited to help save hundreds of millions of dollars. In all of the aforementioned studies, the effect of the options bought with the vehicles was not considered.

A study concerning solution methods is one by Bastin, Cirillo and Toint (2003) [6], where they proposed an algorithm in order to solve mixed logit models. This algorithm adaptively changed the integration sample sizes based on the integration sample error and simulation bias. The sampling was only done for the integration samples, not for the population of individuals. Moreover, the sample sizes were the
same for each individual. For this algorithm, Bastin, Cirillo and Toint (2006) [7] presented convergence results for the parameter estimates when the solutions to the sample average approximation problem were local or first-order critical. Their results covered nonconcave objective functions. This algorithm was applied to simulated data and a real case study in the context of a recent Belgian transportation model, and its performance was benchmarked against existing tools.
CHAPTER II

ALGORITHM

2.1 Motivation

The motivation for this study is to use mixed logit models to forecast demand by approximating the choice probabilities of individuals. The optimization problem that is to be solved for this purpose is given in (1.6.9). This is an optimization problem with an objective function having multi-dimensional integrals. Let $\theta^*$ denote the optimum solution of this maximum likelihood problem:

$$
\theta^* := \arg \max_{\theta} \left\{ LL(\theta) := \frac{1}{I} \sum_{i=1}^{I} \log \left( \int L_{ij}(\beta)f(\beta|\theta) \, d\beta \right) \right\}.
$$

(2.1.1)

assuming that this solution exists and is unique.

This problem will most of the time be hard due to the multi-dimensional integrals in its objective. This necessitates an approximation, which is shown in (1.6.12). Hence an approximate solution, denoted as $\hat{\theta}^S$, will be generated:

$$
\hat{\theta}^S \in \arg \max_{\theta} \left\{ SLL^S(\theta) := \frac{1}{I} \sum_{i=1}^{I} \log \left( SP_{ij}^S(\theta) \right) \right\}.
$$

(2.1.2)

The approximate solution $\hat{\theta}^S$ depends on the integration sample $S = \{S_1, S_2, \ldots, S_I\}$, and as the size of the integration sample increases, the approximate solution converges to the actual solution, $\theta^*$. By a Uniform Law of Large Numbers it is known that as $\min\{|S_1|, \ldots, |S_I|\}$ tends to infinity, $SP_{ij}^S(\theta)$ converges uniformly in $\theta \in \Theta$ on any compact set $\Theta$ to $P_{ij}(\theta)$ with probability one (w.p.1). It follows that the right hand side of (2.1.2) converges w.p.1 to the right hand side of (2.1.1) uniformly in $\theta \in \Theta$ as the sample sizes $|S_i|, i = 1, \ldots, I$ tend to infinity, provided that the right hand side of (2.1.1) is finite valued for all $\theta \in \Theta$. Accordingly, if the sequence of estimates $\hat{\theta}^S$ remains bounded as $\min\{|S_1|, \ldots, |S_I|\} \to \infty$, then w.p.1 $\hat{\theta}^S$ converges
to $\boldsymbol{\theta}^*$ as $\min\{|S_1|, \ldots, |S_I|\} \to \infty$ (see, e.g., Section 6.2 of Rubinstein and Shapiro (1993) [35], for details).

There are different sampling approaches to approximate the actual problem. One way to accomplish the approximation is to use the same integration sample $S_i$ for each individual $i$, i.e. $S_i = S$, $\forall i$. Another approach is to use independent samples of the same size for each individual, i.e. the integration samples for different individuals are independent. The latter method has been observed to perform better than the former one. This can be explained by the observation that the independent sampling method, as compared with the method of using the same sample, reduces the variance of the estimator given in the right hand side of (2.1.2). Indeed, in the method of using the same sample, the terms $\log(SP_{ij}(\boldsymbol{\theta}))$ are highly positively correlated with each other, which adds additional positive terms to the overall variance of the estimator in the right hand side of (2.1.2) as compared with independent sampling.

The approach developed in this study is to use independent integration samples that do not necessarily have the same cardinality for each individual. Hence, there will be independent integration samples, $S_i$, for each individual $i$ with possibly different sizes. This method requires an extra control mechanism to change the sizes of the integration samples for each individual. This control mechanism will be discussed in Section 2.5. In this new approach, the average integration sample size, $\bar{S} = \frac{1}{T} \sum_{i=1}^{I} |S_i|$, will be proportional to the computation time for each optimization step.

In most cases, the cardinality of the set of individuals, denoted by $\{1, 2, \ldots, I\}$, can be quite large. This suggests that sampling from this set of individuals may help any optimization algorithm in terms of computational speed. One idea that has been implemented in the proposed algorithm is to use a sample of individuals, $N \subseteq \{1, 2, \ldots, I\}$, to approximate the simulated log-likelihood function and to adaptively enlarge this sample until $|N| = I$.

Based on the above, both the population of individuals and the distribution of $\beta$
are sampled to approximate the log-likelihood function \(2.1.1\). Using the integration (simulation) sample \(S\), and the data sample \(N\), the simulated log-likelihood function can be rewritten as:

\[
SLL^S_N(\theta) := \frac{1}{|N|} \sum_{i \in N} \log \left( SP^S_{ij}(\theta) \right).
\]  

(2.1.3)

Then the maximum simulated log-likelihood problem becomes:

\[
\hat{\theta}^S_N \in \arg \max_{\theta} SLL^S_N(\theta).
\]

(2.1.4)

The proposed algorithm uses the data samples to speed-up the iterations. Each function evaluation is directly proportional to the size of the data sample. In order for the algorithm to stop, it is required that the final data sample cover the entire dataset of individuals, \(N = \{1, 2, \ldots, I\}\), in which case the subscript \(N\) will be dropped from the notation of the simulated log-likelihood function \(SLL^S(\theta)\), gradient vector \((g^S(\theta))\), the Hessian matrix \((H^S(\theta))\), and the approximate solution returned by the algorithm \(\hat{\theta}^S\). Accordingly, the notation \(\hat{\theta}^S\) implies that no more sampling is done from the dataset of individuals and \(N = \{1, 2, \ldots, I\}\) when the algorithm returns an approximate solution \(\hat{\theta}^S\).

The quality of the approximate solution \(\hat{\theta}^S\) obtained from \(2.1.2\) depends on how well the integral is represented by the integration sample \(S\). Thus, there is a need to statistically test the quality of the solution \(\hat{\theta}^S\). According to our knowledge, there is no algorithm that incorporates such statistical testing within the optimization problem.

The algorithm proposed in this study is a new trust-region based algorithm that solves mixed logit problems by maximizing the simulated log-likelihood function in \(2.1.4\), which furthermore employs statistical testing for the quality of the generated solutions. The algorithm adaptively changes the sizes of both the data sample \(N\) and the integration sample \(S\). The latter enables the algorithm to stop when the approximate solution is satisfactory, meaning that the integration sample is a good
representation of the integral. Trust-region algorithms are reviewed in the following section.

2.2 Trust-Region Algorithms

Trust-region algorithms are popular methods for solving unconstrained optimization problems. They are iterative algorithms that typically use a quadratic approximation of the objective function at each iteration called a model function. This model function is close to the true objective in a region around the current point, and this region is called the trust-region. An important feature of trust-region algorithms is that the size of the trust-region is adaptively changed.

Consider an unconstrained optimization problem with objective function $f$:

$$\max_{x \in \mathbb{R}^n} f(x)$$ \hspace{1cm} (2.2.1)

At iteration $t$, the current point is denoted by $x_t$, and at this point, the function value, the gradient vector and the Hessian matrix are represented by $f(x_t)$, $\nabla f(x_t)$ and $\nabla^2 f(x_t)$, respectively. Let $H_t$ denote an approximation of $\nabla^2 f(x_t)$.

Furthermore, the trust-region centered at $x_t$ is denoted by $TR_t$. Any candidate point can be represented by $x_t + p$, where $p$ is the step taken from point $x_t$ at iteration $t$:

$$(x_t + p) \in TR_t$$

In this study, the trust-region $TR_t$ will be assumed to consist of a ball of radius $r_t$ centered at $x_t$. Hence, the above condition can be rewritten as:

$$\|p\| \leq r_t \iff x_t + p \in TR_t$$

Then, using a quadratic approximation, the model function $m_t$ is constructed as follows:

$$m_t(p) := f(x_t) + \nabla f(x_t)^t p + \frac{1}{2} p^t H_t p$$
The candidate point is found by solving the following subproblem (exactly or approximately):

$$\max_{||p|| \leq r_t} m_t(p)$$  \hspace{1cm} (2.2.2)

The above subproblem is a quadratic problem with quadratic constraints. In order to obtain convergence and good practical behavior, just an approximate solution to the subproblem is sufficient.

As mentioned before, the size of the trust-region is adaptively changed. For this purpose, the so-called *match* between the model function and the objective function is used. Given a step $p_t$, the ratio between the objective function improvement and the model function improvement is used as match:

$$\rho_t := \frac{f(x_t + p_t) - f(x_t)}{m(p_t) - m(0)}$$ \hspace{1cm} (2.2.3)

In this ratio, the denominator, which is the predicted improvement, is always non-negative. If the ratio $\rho_t$ is close to 1, the match between the model function and the objective function is good and this promotes the intuition to grow the size of the trust-region so that a larger (longer) step can be taken in the next iteration. If on the other hand, this ratio is negative or small, the match is not satisfactory and the trust-region should be shrunk in order to make the match better. The classical trust-region algorithm is presented by means of its pseudocode in the next section.

### 2.2.1 Trust-Region Algorithm Pseudocode

The pseudocode of the trust-region algorithm as presented by Nocedal and Wright (1999) \[30\] is given in Algorithm 1.

### 2.2.2 Approximate Solution of the Subproblems

Ideally, the candidate solution $x_t + p_t$ is calculated by solving the subproblem (2.2.2). For the convergence of trust-region algorithms it is not necessary to solve the subproblem exactly. In this section, three methods that solve the subproblem approximately
Algorithm 1 Trust-Region Algorithm
1: Given $\bar{r} > 0$, $r_0 \in (0, \bar{r})$, and $\eta \in [0, 1/4)$:
2: Let $H_0$ be the initial Hessian approximation.
3: for $t = 0, 1, \ldots$ do
4: Obtain $p_t$ by (approximately) solving (2.2.2)
5: Update the approximate Hessian $H_t$ in (2.2.2) by an update method. See Section 2.4.2 for details.
6: Evaluate $\rho_t$ from (2.2.3)
7: if $\rho_t < \frac{1}{4}$ then
8: $r_{t+1} = \frac{1}{4}||p_t||$
9: else
10: if $\rho_t > \frac{3}{4}$ and $||p_t|| = r_t$ then
11: $r_{t+1} = \min(2r_t, \bar{r})$
12: else
13: $r_{t+1} = r_t$
14: end if
15: end if
16: if $\rho_t > \eta$ then
17: $x_{t+1} = x_t + p_t$
18: else
19: $x_{t+1} = x_t$
20: end if
21: end for
will be discussed. These methods are the Cauchy Point Method, the Dogleg Method
and the Steihaug Method.

2.2.2.1 Cauchy Point Method

The Cauchy Point Method is one of the easiest methods to approximately solve the
subproblem (2.2.2). The method makes a sufficient improvement in the subproblem
such that eventually the convergence of the actual problem (2.2.1) is guaranteed, and
the proof is given in Nocedal and Wright (1999) [30]. The pseudocode of the Cauchy
Point Method is given in Algorithm 2.

Algorithm 2 Cauchy Point Method
1: Calculate the point \( p^*_t \) that solves the linear version of the subproblem:

\[
p^*_t = \arg \max_{||p|| \leq \tau_t} f(x_t) + \nabla f(x_t)p.
\]

2: Calculate the scalar \( \tau_t > 0 \) such that:

\[
\tau_t = \arg \max_{0 \leq \tau \leq 1} m_t(\tau p^*_t).
\]

3: The Cauchy point \( p^*_i = \tau_t p^*_t \).

The other two approximate solution methods, namely the Dogleg and Steihaug
methods, basically improve on the Cauchy point by actually originating on the Cauchy
point and then considering a better point using the Hessian or the approximation of
the Hessian.

2.2.2.2 Dogleg Method

This method is applicable for problems having objective functions with a negative
definite Hessian or Hessian approximation. The method first calculates the uncon-
strained maximizer \( p^B \) of the model function. If the unconstrained maximizer \( p^B \) is
within the trust-region, then it also becomes the constrained maximizer of the model
function. Otherwise, the method calculates the unconstrained maximizer along the
steepest ascent direction, \( p^U \). This allows the method to construct a trajectory from
the current point $x_t$ to $x_t + p^U$ and from there to $x_t + p^B$. The shape of the trajectory gives its name, Dogleg. The pseudocode of the dogleg method as presented in Nocedal and Wright (1999) [30] is given in Algorithm 3.

**Algorithm 3 Dogleg Method**

1. Calculate the unconstrained maximizer $p^B$ of the model function $m_t(p)$:
   \[ p^B := -H_t^{-1}\nabla f(x_t). \]

2. if $||p^B|| \leq r_t$ then

4. else
   5. Calculate the unconstrained maximizer along the steepest ascent direction, $p^U$:
   \[ p^U := -\frac{\nabla f(x_t)\nabla f(x_t)}{\nabla f(x_t)'H_t\nabla f(x_t)}\nabla f(x_t). \]

6. Define the dogleg trajectory:
   \[ \tilde{p}(\tau) := \begin{cases} \tau p^U, & 0 \leq \tau \leq 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 2. \end{cases} \]

7. Return $\tilde{p}(\tau^*)$, which intersects the trust-region boundary.
8. end if

2.2.2.3 Steihaug Method

The Dogleg method discussed in the previous section requires the inverse of the Hessian, which is computationally challenging if the size of the matrix is large. The Steihaug method on the other hand, generates an approximate solution based on the conjugate gradient algorithm and an improvement over the Cauchy Point method. This method can also handle Hessian matrices that are not positive definite. The pseudocode of the Steihaug method as presented in Steihaug (1983) [38] is given in Algorithm 4.
\textbf{Algorithm 4 Steihaug Method}

1: Given $\varepsilon > 0$. Set $p_0 = 0$, $q_0 = \nabla f(x_0)$, $d_0 = q_0$;
2: \hspace{1em} for $j = 0, 1, \ldots$ do
3: \hspace{2em} Compute $\gamma_j = d'_j H_j d_j$
4: \hspace{2em} if $\gamma_j \geq 0$ then
5: \hspace{3em} Find $\tau > 0$ such that $||p_j + \tau d_j|| = r_t$
6: \hspace{3em} Return $p = p_j + \tau d_j$.
7: \hspace{2em} end if
8: \hspace{2em} Set $\alpha_j = \frac{r_j r_{j+1}}{\gamma_j}$;
9: \hspace{2em} Set $p_{j+1} = p_j + \alpha_j d_j$;
10: \hspace{2em} if $||p_{j+1}|| \geq r_t$ then
11: \hspace{3em} Find $\tau$ such that $p = p_j + \tau d_j$ satisfies $||p|| = r_t$;
12: \hspace{3em} Return $p$.
13: \hspace{2em} end if
14: \hspace{2em} Set $q_{j+1} = q_j - \alpha_j H_j d_j$;
15: \hspace{2em} if $||q_{j+1}|| < \varepsilon ||\nabla f(x_0)||$ then
16: \hspace{3em} Return $p = p_{j+1}$.
17: \hspace{2em} end if
18: \hspace{2em} $\beta_j = \frac{q_{j+1} d_{j+1}}{q_j d_j}$;
19: \hspace{2em} $d_{j+1} = q_{j+1} + \beta_j d_j$;
20: \hspace{2em} end for

\subsection{2.3 Errors}

As explained in Section 2.1, the maximum log-likelihood problem attempted in this study is approximated by the use of two types of samples; the data sample $(N)$ and the integration sample $(S)$. Accordingly, for both of these samples, there are respective errors introduced to the approximation. These errors are important in that at each iteration of the proposed algorithm, they will be used in the decision mechanism for the progress of the algorithm, as will be explained in more detail in section 2.4.

A quick overview to the role played by the error terms is as follows. Since the approximate log-likelihood problem in (2.1.4) is maximized over $\theta \in \mathbb{R}^K$ (namely the distribution parameters of $\beta$), given two candidate values for the parameter $\theta$, namely $\theta_1$ and $\theta_2$, the log-likelihood function values at these points are represented by $LL(\theta_1)$ and $LL(\theta_2)$, respectively. The difference between these log-likelihood function values
is denoted by $\Delta LL(\theta_1, \theta_2)$:

$$\Delta LL := LL(\theta_1) - LL(\theta_2). \quad (2.3.1)$$

Because of the high dimensional integrals in the log-likelihood functions, it is in most cases impossible to calculate the log-likelihood function values exactly, for which reason the integration sample $S$ is adopted. As a result of this, the integration sample error is introduced. On top of that, the dataset is approximated by the data sample $(N)$, which further introduces the data sample error. Therefore the simulated log-likelihood function values are used to estimate $(2.3.1)$:

$$\Delta SLL := SLL^S_N(\theta_1) - SLL^S_N(\theta_2). \quad (2.3.2)$$

By using the difference in the simulated log-likelihood function values instead of the log-likelihood function values themselves, two types of errors are incurred.

For a fixed integration sample $S = \{S_i, i \in N\}$, denote the error due to sampling from the dataset of individuals as $\sigma^2_N(\theta_1, \theta_2)$ (See Section 2.3.1 for details):

$$\sigma^2_N(\theta_1, \theta_2) := \text{Var}_N(SLL^S_N(\theta_1) - SLL^S_N(\theta_2) \mid S)$$

$$= \text{Var} \left( \frac{1}{|N|} \sum_{i \in N} \{ \log(SP^S_{i,j}(\theta_1)) - \log(SP^S_{i,j}(\theta_2)) \} \mid S \right) \quad (2.3.3)$$

For a fixed data sample $N$, denote the error due to sampling for the multidimensional integrals (integration sample error) as $\sigma^2_S(\theta_1, \theta_2)$ (See Section 2.3.2 for details):

$$\sigma^2_S(\theta_1, \theta_2) := \text{Var}_S(SLL^S_N(\theta_1) - SLL^S_N(\theta_2) \mid N)$$

$$= \text{Var} \left( \frac{1}{|N|} \sum_{i \in N} \{ \log(SP^S_{i,j}(\theta_1)) - \log(SP^S_{i,j}(\theta_2)) \} \mid N \right) \quad (2.3.4)$$

The algorithm uses both the data sample error $\sigma^2_N(\theta_1, \theta_2)$ and the integration sample error $\sigma^2_S(\theta_1, \theta_2)$ to manipulate the integration samples for each individual and the data sample. However, these error terms are also hard to calculate and therefore need to be estimated.
In addition to the above, another critical element of the decision mechanism of the algorithm is the gradient of the log-likelihood function, \( \nabla LL(\theta) \), which should be used in defining the stopping condition of the algorithm. One of the stopping criteria to use in such maximum log-likelihood problems is that \( ||\nabla LL(\theta)|| \) is close to zero.

As mentioned before, it is most of the time impossible to calculate \( LL(\theta) \) exactly, and this also applies for \( \nabla LL(\theta) \). Hence, \( \nabla LL(\theta) \) should be approximated by \( g^S(\theta) \), as given in Appendix (C.2.2). However, approximating \( \nabla LL(\theta) \) using the integration sample \( S \) introduces error to the calculated value. Thus, \( ||g^S(\theta)||^2 \approx 0 \) is not a good stopping criterion due to the implicit error caused by the integration sample \( S \). The stopping criterion must take account of the error implicit in the gradient \( g^S(\theta) \) and this is done by considering the variance of \( ||g^S(\theta)||^2 \) with respect to the integration sample \( S \). This is shown in more detail in Section 2.3.3.

### 2.3.1 Data Sample Error

For a fixed integration sample \( S = \{S_i, \ i \in N\} \), variance of the difference in the simulated log-likelihood function values for the parameters \( \theta_1 \) and \( \theta_2 \) with respect to data sample \( N \) is denoted by \( \sigma^2_N(\theta_1, \theta_2) \). Given the definition of simulated log-likelihood function in (2.1.3):

\[
\sigma^2_N(\theta_1, \theta_2) = \text{Var} \left( \frac{1}{|N|} \sum_{i \in N} \{ \log \left( SP^S_{ij}(\theta_1) \right) - \log \left( SP^S_{ij}(\theta_2) \right) \} \ | \ S \right).
\]

The sample of individuals \( N \) is obtained from a finite population \( I \) and the sampling is done with equal probabilities without replacement. Thus a correction factor for finite populations, denoted as \( CF \), is used.

\[
\sigma^2_N(\theta_1, \theta_2) = CF \times \text{Var} \left( \frac{1}{|N|} \sum_{i \in N} \{ \log \left( SP^S_{ij}(\theta_1) \right) - \log \left( SP^S_{ij}(\theta_2) \right) \} \ | \ S \right),
\]

where

\[
CF := \left( \frac{I - |N|}{I - 1} \right).
\]
The data sample error \( \sigma^2_N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \) can be approximated by:

\[
\sigma^2_N(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \approx \frac{CF}{|N|(|N|-1)} \times \sum_{i \in N} \{ \log (SP_{ij}^S(\boldsymbol{\theta}_1)) - \log (SP_{ij}^S(\boldsymbol{\theta}_2)) \} - \left( \frac{1}{|N|} \sum_{i \in N} \{ \log (SP_{ij}^S(\boldsymbol{\theta}_1)) - \log (SP_{ij}^S(\boldsymbol{\theta}_2)) \} \right)^2.
\]

### 2.3.2 Integration (Simulation) Sample Error

For a fixed data sample \( N \), the variance of the difference in the simulated log-likelihood function values for the parameters \( \boldsymbol{\theta}_1 \) and \( \boldsymbol{\theta}_2 \) with respect to the integration sample \( S \) is denoted by \( \sigma^2_S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \). Using the fact that independent integration samples are used for each individual \((S_i, i \in N)\), \( \sigma^2_S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \) can be calculated as follows:

\[
\sigma^2_S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \text{Var}_S \left( \frac{1}{|N|} \sum_{i \in N} \{ \log (SP_{ij}^S(\boldsymbol{\theta}_1)) - \log (SP_{ij}^S(\boldsymbol{\theta}_2)) \} \bigg| N \right)
= \frac{1}{|N|^2} \sum_{i \in N} \text{Var}_S \{ \log (SP_{ij}^S(\boldsymbol{\theta}_1)) - \log (SP_{ij}^S(\boldsymbol{\theta}_2)) \}.
\]

The terms inside the variance operator can be approximated by the first order Taylor series around the actual (unknown) choice probabilities \( P_{ij}(\boldsymbol{\theta}_1) \) and \( P_{ij}(\boldsymbol{\theta}_2) \):

\[
\log(SP_{ij}^S(\boldsymbol{\theta}_1)) \approx \log (P_{ij}(\boldsymbol{\theta}_1)) + \frac{1}{P_{ij}(\boldsymbol{\theta}_1)} (SP_{ij}^S(\boldsymbol{\theta}_1) - P_{ij}(\boldsymbol{\theta}_1)) \]
\[
\log(SP_{ij}^S(\boldsymbol{\theta}_2)) \approx \log (P_{ij}(\boldsymbol{\theta}_2)) + \frac{1}{P_{ij}(\boldsymbol{\theta}_2)} (SP_{ij}^S(\boldsymbol{\theta}_2) - P_{ij}(\boldsymbol{\theta}_2)).
\]

Using these approximations, the integration error \( \sigma^2_S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \) can be estimated by:

\[
\sigma^2_S(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \approx \frac{1}{|N|^2} \sum_{i \in N} \{ \text{Var}_S \left( \log (P_{ij}(\boldsymbol{\theta}_1)) + \frac{1}{P_{ij}(\boldsymbol{\theta}_1)} (SP_{ij}^S(\boldsymbol{\theta}_1) - P_{ij}(\boldsymbol{\theta}_1)) \right) - \log (P_{ij}(\boldsymbol{\theta}_2)) - \frac{1}{P_{ij}(\boldsymbol{\theta}_2)} (SP_{ij}^S(\boldsymbol{\theta}_2) - P_{ij}(\boldsymbol{\theta}_2)) \}.
\]

The terms \( \log (P_{ij}(\boldsymbol{\theta}_1)) \) and \( \log (P_{ij}(\boldsymbol{\theta}_2)) \) in the variance operator are constants,
which yields:

\[
\sigma^2_s(\theta_1, \theta_2) \approx \frac{1}{|N|^2} \sum_{i \in N} \text{Var}_S \left( \frac{1}{|S_i|} \sum_{\nu \in S_i} \left( \frac{L_{ij}(\beta^\nu(\theta_1))}{P_{ij}(\theta_1)} - \frac{L_{ij}(\beta^\nu(\theta_2))}{P_{ij}(\theta_2)} \right) \right)
\]

\[
\approx \frac{1}{|N|^2} \sum_{i \in N} \frac{1}{|S_i|(|S_i| - 1)} \sum_{\nu \in S_i} \left( \frac{L_{ij}(\beta^\nu(\theta_1))}{SP_{ij}(\theta_1)} - \frac{L_{ij}(\beta^\nu(\theta_2))}{SP_{ij}(\theta_2)} \right) \right) \right)^2.
\]

### 2.3.3 Error in Gradient Approximation

Let \( \sigma^2_{|g^s(\theta)|^2} \) be the integration sample variance of \( |g^s(\theta)|^2 \) at the parameter \( \theta \). Note that \( \sigma^2_{|g^s(\theta)|^2} \) is only calculated if \( |N| = I \), i.e., all the individuals in the population are selected for the data sample \( N \).

Using the definition of \( g^s(\theta) \) in equation (C.2.2), and the fact that independent integration samples are used for each individual \( (S_i, i \in N) \):

\[
\sigma^2_{|g^s(\theta)|^2} := \text{Var} \left( \left( \frac{1}{I} \sum_{i=1}^I \frac{1}{SP_{ij}^S(\theta)} \nabla SP_{ij}^S(\theta) \right) \right) \left( \left( \frac{1}{I} \sum_{i=1}^I \frac{1}{SP_{ij}^S(\theta)} \nabla SP_{ij}^S(\theta) \right) \right)
\]

\[
= \frac{1}{I^4} \sum_{i=1}^I \text{Var} \left( \frac{1}{(SP_{ij}^S(\theta))^2} \left( \nabla SP_{ij}^S(\theta) \right)' \nabla SP_{ij}^S(\theta) \right)
\]

\[
+ \frac{4}{I^4} \sum_{i<k} \text{Var} \left( \frac{1}{SP_{ij}^S(\theta)} \frac{1}{SP_{ik}^S(\theta)} \left( \nabla SP_{ij}^S(\theta) \right)' \nabla SP_{ij}^S(\theta) \right).
\]

(2.3.5)

Let

\[
h_{ii}(SP_{ij}^S(\theta), \nabla SP_{ij}^S(\theta)) := \frac{1}{(SP_{ij}^S(\theta))^2} \left( \nabla SP_{ij}^S(\theta) \right)' \nabla SP_{ij}^S(\theta)
\]

denote the term inside the first variance operator on the right hand side of (2.3.5).
This function \( h_{ii}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta)) \) can be approximated by the first order Taylor approximation:

\[
h_{ii}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta)) \approx h_{ii}(P_{ij}(\theta), \nabla P_{ij}(\theta)) \\
+ \nabla h_{ii}(P_{ij}(\theta), \nabla P_{ij}(\theta))' \begin{pmatrix} SP_{ij}^i(\theta) - P_{ij}(\theta) \\ \nabla SP_{ij}^i(\theta) - \nabla P_{ij}(\theta) \end{pmatrix},
\]

where

\[
\nabla h_{ii}(P_{ij}(\theta), \nabla P_{ij}(\theta)) = \begin{pmatrix} \frac{2}{(P_{ij}(\theta))^3} \nabla P_{ij}(\theta)' \nabla P_{ij}(\theta) \\ \frac{2}{(P_{ij}(\theta))^2} \nabla P_{ij}(\theta) \end{pmatrix}.
\]

Using this approximation, the first variance operator in (2.3.5) can be estimated by:

\[
\text{Var}(h_{ii}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta))) \approx \nabla h_{ii}(P_{ij}(\theta), \nabla P_{ij}(\theta))' \text{Cov} \begin{pmatrix} SP_{ij}^i(\theta) - P_{ij}(\theta) \\ \nabla SP_{ij}^i(\theta) - \nabla P_{ij}(\theta) \end{pmatrix} \nabla h_{ii}(P_{ij}(\theta), \nabla P_{ij}(\theta)).
\]

The terms inside the second variance operator on the right hand side of (2.3.5) can be rewritten as:

\[
h_{ik}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta), SP_{kj}^k(\theta), \nabla SP_{kj}^k(\theta)) := \frac{1}{SP_{ij}^i(\theta) SP_{kj}^k(\theta)} (\nabla SP_{ij}^i(\theta))' \nabla SP_{kj}^k(\theta).
\]

For \( i < k \), this function \( h_{ik}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta), SP_{kj}^k(\theta), \nabla SP_{kj}^k(\theta)) \) can be approximated by the first order Taylor approximation:

\[
h_{ik}(SP_{ij}^i(\theta), \nabla SP_{ij}^i(\theta), SP_{kj}^k(\theta), \nabla SP_{kj}^k(\theta)) \\
\approx h_{ik}(P_{ij}(\theta), \nabla P_{ij}(\theta), P_{kj}(\theta), \nabla P_{kj}(\theta)) \\
+ \nabla h_{ik}(P_{ij}(\theta), \nabla P_{ij}(\theta), P_{kj}(\theta), \nabla P_{kj}(\theta))' \begin{pmatrix} SP_{ij}^i(\theta) - P_{ij}(\theta) \\ \nabla SP_{ij}^i(\theta) - \nabla P_{ij}(\theta) \\ SP_{kj}^k(\theta) - P_{kj}(\theta) \\ \nabla SP_{kj}^k(\theta) - \nabla P_{kj}(\theta) \end{pmatrix},
\]
where

\[
\nabla h_{ik}(SP_{ij}(\theta), \nabla SP_{ij}(\theta), SP_{kjh}(\theta), \nabla SP_{kjh}(\theta)) = 
\begin{pmatrix}
-\frac{1}{(P_{ij}(\theta))^2} \frac{1}{P_{kjh}(\theta)} \nabla P_{ij}(\theta) \nabla P_{kjh}(\theta) \\
\frac{1}{P_{ij}(\theta)} \frac{1}{P_{kjh}(\theta)} \nabla P_{ij}(\theta) \\
-\frac{1}{P_{ij}(\theta)} \frac{1}{P_{kjh}(\theta)} \nabla P_{ij}(\theta) \nabla P_{kjh}(\theta) \\
\frac{1}{P_{ij}(\theta)} \frac{1}{P_{kjh}(\theta)} \nabla P_{ij}(\theta) 
\end{pmatrix}.
\]

Using this approximation, the variance of the function above can be estimated by:

\[
\text{Var}(h_{ik}(SP_{ij}(\theta), \nabla SP_{ij}(\theta), SP_{kjh}(\theta), \nabla SP_{kjh}(\theta)))
\approx \nabla h_{ik}(SP_{ij}(\theta), \nabla SP_{ij}(\theta), SP_{kjh}(\theta), \nabla SP_{kjh}(\theta))'
\begin{pmatrix}
SP_{ij}(\theta) - P_{ij}(\theta) \\
\nabla SP_{ij}(\theta) - \nabla P_{ij}(\theta) \\
SP_{kjh}(\theta) - P_{kjh}(\theta) \\
\nabla SP_{kjh}(\theta) - \nabla P_{kjh}(\theta)
\end{pmatrix}.
\text{Cov}
\]

\[
(2.3.7)
\]

The variance of \(||g^S(\theta)||^2\) given in (2.3.5) can be estimated using the approximations in (2.3.6) and (2.3.7).
2.4 Pseudocode

This section presents the pseudocode of the trust-region based algorithm developed in this study. The pseudocode consists of six steps, of which the details are explained in the following subsections.

2.4.1 Step 0: Initialization

Let \( t \) be the iteration count of the algorithm initialized by \( t = 0 \).

Define constants \( C = 0.5 \), \( \eta = 0.05 \) and \( K_{ErrTol} = 10^{-9} \).

Generate initial data sample \( N_0 \subseteq \{1, 2, \ldots, I\} \). This is accomplished by randomly selecting the individuals from the entire population \( \{1, 2, \ldots, I\} \). Let \( p_0^N = N_0/I \) be the probability that any individual in the population will be selected for the initial data sample \( N_0 \).

Initialize starting point \( \theta_0 \), trust-region radius \( r_0 \).

Generate initial integration sample \( S_0 = \{S_{0,i} \mid i \in N_0\} \). So for every individual \( i \) selected to be in the data sample, an initial independent integration sample of size \( |S_{0,i}| \) is constructed.

Calculate the simulated log-likelihood \( SLL_{N_0}^{S_0}(\theta_0) \), the gradient \( g_{N_0}^{S_0}(\theta_0) \), and the Hessian \( H_{N_0}^{S_0}(\theta_0) \) functions. These functions are given in Appendix (C.2.1), Appendix (C.2.2) and Appendix (C.2.6), respectively. If a Hessian approximation method (Quasi Newton method) is used, then the Hessian matrix \( H_{N_0}^{S_0}(\theta_0) \) is not calculated here, instead the Hessian matrix is initialized to the negative Identity matrix.

2.4.2 Step 1: Model Function Optimization

Suppose that at iteration \( t \), \( \theta_t \) is the current parameter vector and the proposed algorithm uses \( N_t \) and \( S_t \) as the data sample and the integration sample, respectively. The objective function value, in other words the simulated log-likelihood function value at iteration \( t \) using samples \( N_t \) and \( S_t \), denoted by \( SLL_{N_t}^{S_t}(\theta_t) \), is given in (2.1.3). In
order to construct the Taylor approximation, the gradient and a Hessian of the simulated log-likelihood function are required. For the calculation of the Hessian, there are two options, namely to calculate either the actual Hessian or the approximate Hessian. If the latter is selected, a Quasi Newton method is used for the Hessian approximation. This approximate Hessian is updated at the end of Step 1 at each iteration. The gradient and the Hessian (or approximate Hessian) of the simulated log-likelihood function are denoted by $g^S_{N_t}(\theta_t)$ and $H^S_{N_t}(\theta_t)$, respectively.

Hence, at the current parameter vector $\theta_t$, the trust-region model function can be constructed as:

$$m^S_{N_t}(\theta) := SLL^S_{N_t}(\theta_t) + g^S_{N_t}(\theta_t)'(\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)'H^S_{N_t}(\theta_t)(\theta - \theta_t). \quad (2.4.1)$$

At this iteration, using the trust-region radius $r_t$, the trust-region $TR_t$ is as follows:

$$TR_t := \{\theta \mid ||\theta - \theta_t|| \leq r_t\}. \quad (2.4.2)$$

The trust-region is the region over which the model function is believed to be a good approximation of the actual function, which is the simulated log-likelihood function in this case. Solving the maximization problem for this model function over the trust-region $TR_t$ instead of solving the original function is a much easier task. Hence, the subproblem becomes:

$$\theta^*_t \in \arg \max_{\theta \in TR_t} m^S_{N_t}(\theta). \quad (2.4.3)$$

In the above subproblem, $\theta^*_t$ is the candidate point generated at iteration $t$, which can be accepted to become the next parameter vector, $\theta_{t+1}$, or rejected. This subproblem is a concave quadratic problem. For the local convergence of trust-region algorithms in general, discussed by Nocedal and Wright (1999) [30], it is desirable that $\theta^*_t$ increases the model function $m^S_{N_t}$ at least as much as the Cauchy Point Method. As discussed in Section 2.2.2, three methods available in the literature have been implemented in this study for the approximate solution of the subproblems:
• Cauchy Point Method (See Section 2.2.2.1)

• Dogleg Method (See Section 2.2.2.2)

• Steihaug Method (See Section 2.2.2.3)

The quality of the approximation within the trust-region can be tested by the ratio \( \rho_t \), which is the ratio of the improvement in the simulated log-likelihood function value to the predicted improvement in the model function. This ratio is an indication of whether the match between the model function and the objective function is satisfactory or not. The predicted improvement in the simulated log-likelihood function, denoted by \( \Delta SLL_{N_t}(\theta_t, \theta_t^e) \), is:

\[
\Delta SLL_{N_t}(\theta_t, \theta_t^e) := SLL_{N_t}(\theta_t^e) - SLL_{N_t}(\theta_t).
\]

The predicted improvement in the model function, denoted by \( \Delta m_{N_t}(\theta_t, \theta_t^e) \), is:

\[
\Delta m_{N_t}(\theta_t, \theta_t^e) := m_{N_t}(\theta_t^e) - m_{N_t}(\theta_t).
\]

The size of the trust-region is adaptively changed based on the ratio \( \rho_t \) as will be explained in Section 2.4.6:

\[
\rho_t := \frac{\Delta SLL_{N_t}(\theta_t, \theta_t^e)}{\Delta m_{N_t}(\theta_t, \theta_t^e)}.
\]

The task of updating the approximate Hessian depends on the type of Hessian update method selected. Two different Quasi-Newton methods can be used, namely the BFGS and SR1 methods. The Hessian update operations required for these methods are given below:

• BFGS method

\[
H_{N_{t+1}}^{S_{t+1}}(\theta_{t+1}) = H_{N_t}^{S_t}(\theta_t) - \frac{H_{N_t}^{S_t}(\theta_t)s_t}{s_t' H_{N_t}^{S_t}(\theta_t)s_t} + \frac{y_t'y_t}{y_t's_t},
\]

where

\[
y_t = g_{N_t}^{S_t}(\theta_t^e) - g_{N_t}^{S_t}(\theta_t), \text{ and } s_t = \theta_t^e - \theta_t.
\]
• SR1 method

\[
H_{N_t+1}^{S_t+1}(\theta_{t+1}) = H_{N_t}^S(\theta_t) + \frac{(s_t - H_{N_t}^S(\theta_t)y_t)(s_t - H_{N_t}^S(\theta_t)y_t)^\prime}{(s_t - H_{N_t}^S(\theta_t)y_t)^\prime y_t}.
\]

Hence, Step 1 of the pseudocode is given in Algorithm 5.

**Algorithm 5 Model Function Optimization**

1: Calculate the candidate point \( \theta_t^c \) in (2.4.3) as explained in Section 2.2.2.3.
2: At \( \theta_t^c \), calculate:
   • the simulated log-likelihood function value \( SLL_{N_t}^S(\theta_t^c) \)
   • the gradient of the simulated log-likelihood function \( g_{N_t}^S(\theta_t^c) \)
   • the Hessian of the simulated log-likelihood function \( H_{N_t}^S(\theta_t^c) \) (if the actual Hessian is used)
3: Using \( SLL_{N_t}^S(\theta_t^c) \), \( g_{N_t}^S(\theta_t^c) \), and \( H_{N_t}^S(\theta_t^c) \) calculate:
   • the predicted improvement in the model function:
     \[
     \Delta m_{N_t}^S(\theta_t, \theta_t^c) := g_{N_t}^S(\theta_t)^\prime (\theta_t^c - \theta_t) + \frac{1}{2}(\theta_t^c - \theta_t)^\prime H_{N_t}^S(\theta_t)(\theta_t^c - \theta_t).
     \]
   • the increase in the simulated log-likelihood function:
     \[
     \Delta SLL_{N_t}^S(\theta_t, \theta_t^c) := SLL_{N_t}^S(\theta_t^c) - SLL_{N_t}^S(\theta_t).
     \]
4: Calculate the ratio \( \rho_t \):

\[
\rho_t := \frac{\Delta SLL_{N_t}^S(\theta_t, \theta_t^c)}{\Delta m_{N_t}^S(\theta_t, \theta_t^c)}
\]
5: If Hessian matrix approximation is selected, then update the Hessian matrix \( H_{N_t+1}^{S_t+1}(\theta_{t+1}) \) using BFGS or SR1 update formulas.

### 2.4.3 Step 2: Error Calculation

A brief introduction to errors and a detailed explanation of their calculations were given previously in Section 2.3. As mentioned before, the errors play an important role in determining how the algorithm will proceed at each iteration. In other words, they are the key elements that influence the decision mechanism of the algorithm.
The reasoning behind this fact and the mechanism by which the errors influence the algorithm will be explained thoroughly in Section 2.4.5.

At this stage, it should be noted that calculating the gradient approximation error $\sigma_{||g^s(\theta^c)||^2}$ is computationally much harder than calculating the data sample error and the integration sample error. Since the gradient approximation error is only used for the stopping condition, it is only necessary to calculate it if all of the following hold:

- the set of all individuals is selected,
- the predicted improvement in the objective value is less than the integration sample error, and
- the integration sample error is small.

The pseudocode for the error calculation step of the algorithm is as shown in Algorithm 6.

**Algorithm 6** Error Calculation

1: Calculate the data sample error, $\sigma_N(\theta, \theta^c)$ and the integration sample error, $\sigma_S(\theta, \theta^c)$.
2: if $|N_1| = 1$ & ($\Delta m^S_{|N_1|}(\theta, \theta^c) < C\sigma_S(\theta, \theta^c)$) & ($\sigma_S(\theta, \theta^c) < \tilde{K}_{\text{ErrTot}}$) then
3: calculate $\sigma_{||g^s(\theta)||^2}$.
4: end if

### 2.4.4 Step 3: Stopping Condition

The pseudocode for the stopping condition of the algorithm is as shown in Algorithm 7.

**Algorithm 7** Stopping Condition

1: if $|N_1| = 1$ & ($||g^s(\theta)||^2 + C\sigma_{||g^s(\theta)||^2} \approx 0$) then
2: Stop the algorithm with solution vector $\theta^c$.
3: end if
2.4.5 Step 4: Sample Control Mechanisms

As shown in Step 1 of the algorithm, i.e. Algorithm 5, at each iteration, the improvement on the simulated log-likelihood function is calculated and the quality of the match between the model function and the objective function is tested. A crucial fact to be taken into account at this stage is that the above mentioned improvement is that obtained on the simulated log-likelihood function, and not the log-likelihood function itself. The calculation concerns a comparison of the improvement obtained on the simulated log-likelihood function and the variance of the improvement at that current point with respect to the data sample and the integration sample. If the variances, i.e. the errors, have significantly large values with respect to the improvement on the simulated log-likelihood function, then it can be concluded that although the simulated log-likelihood function shows an improvement, the actual improvement on the log-likelihood function may be not significant. If, on the other hand, the errors are relatively small in comparison with the improvement in the simulated log-likelihood function, then the conclusion is that the improvement on the log-likelihood function is also significant. These conclusions have some implications on the algorithm. These implications and their consequences are explained next.

If the improvement on the log-likelihood function is considered to be significant, this implies that the data sample and the integration (simulation) sample provide a good representation of the real problem and are of adequate size. If the contrary is true, then the latter does not hold and the data sample size or the integration (simulation) sample size has to be increased. The pseudocode for the sample control mechanisms of the algorithm is as shown in Algorithm 8.
Algorithm 8 Sample Control Mechanisms

1: if \( |N| < I \) \& \( \sigma_N(\theta_t, \theta_t^*) \geq \sigma_S(\theta_1, \theta_2) \) \& \( \Delta m^S_{N_t}(\theta_t, \theta_t^*) < C \sigma_N(\theta_t, \theta_t^*) \) then
2: Expand \( N_{t+1} \) using \( N_t \) such that \( |N_{t+1}| = \left( \frac{C \sigma_N(\theta_t, \theta_t^*)}{\Delta m^S_{N_t}(\theta_t, \theta_t^*)} \right)^2 \).
3: \( \theta_{t+1} \leftarrow \theta_t \).
4: \( t \leftarrow t + 1 \).
5: Calculate \( SLL^S_{N_t}(\theta_t) \), \( g^S_{N_t}(\theta_t) \), \( H^S_{N_t}(\theta_t) \).
6: Go to Step 1.
7: end if
8: if \( \sigma_S(\theta_t, \theta_t^*) > \sigma_N(\theta_t, \theta_t^*) \) \& \( \Delta m^S_{N_t}(\theta_t, \theta_t^*) \leq C \sigma_N(\theta_t, \theta_t^*) \) then
9: Expand \( S_{t+1} \) using the algorithm in Section 2.5.
10: \( \theta_{t+1} \leftarrow \theta_t \).
11: \( t \leftarrow t + 1 \).
12: Calculate \( SLL^S_{N_t}(\theta_t) \), \( g^S_{N_t}(\theta_t) \), \( H^S_{N_t}(\theta_t) \).
13: Go to Step 1.
14: end if

2.4.6 Step 5: Trust-Region Control Mechanism

The classical trust-region control mechanism is adopted, in which the quality of the match between the objective function improvement and the model function improvement is tested by their ratio, \( \rho_t \). If \( \rho_t \) is small, then the approximation can be improved by reducing the trust-region radius \( r_t \). If the ratio is big enough, then it can be concluded that the model function is a good-fit and the trust-region radius can be enlarged so that the algorithm can proceed with larger steps. The pseudocode for the trust-region control mechanism of the algorithm is as shown in Algorithm 9.

2.5 Optimum Allocation of Integration Sample

The relation between the integration sample size \( S \), and the integration sample error \( \sigma_S(\theta_t, \theta_t^*) \) has been analyzed in the former sections and it follows that this error is inversely proportional to the integration sample sizes of the individuals. It has been previously mentioned that one of the aims of the algorithm is also to reduce the error term. Thus, a few features of the algorithm need to be summarized at this stage. As mentioned before, the integration samples for each individual are not only

46
Algorithm 9 Trust-Region Control Mechanism

1: if $\rho_t < 0.10$ then
2: $r_{t+1} = \|s_t\|/2$.
3: else
4: if $\rho_t > 0.80 \& \|\theta_t^e - \theta_t^c\| \approx r_t$ then
5: $r_{t+1} = 2 \times r_t$.
6: end if
7: end if
8: if $\rho_t > \eta$ then
9: $\theta_{t+1} \leftarrow \theta_t^c$
10: else
11: $\theta_{t+1} \leftarrow \theta_t$
12: end if
13: $t \leftarrow t + 1$.

independent but the sizes of the samples may also be different. Accordingly, a single common integration sample size does not exist, for which reason an average of the integration sample sizes is needed. This average integration sample size has also a practical role in determining the running time of the algorithm in that it is directly proportional to the computation time of each iteration.

Another way of explaining the above is that the integration sample sizes need to be increased in order to reduce the integration sample error. This follows automatically from the fact that the error term is inversely proportional to the respective integration sample sizes of the individuals. Increasing the integration sample sizes, however, has the inevitable effect of increasing the computational effort required to compute the function evaluations at each iteration. Bearing in mind that one of the aims is to keep the computational effort at a minimum, conflicting objectives arise, bringing up a practical trade-off problem. What is desired is to allocate the minimum amount of computational effort by imposing the objective that the computational time must be kept at the minimum possible value, while simultaneously reducing the integration error by a predetermined constant factor.

The new integration sample will be denoted by $S' = \{S'_1, S'_2, \ldots, S'_t\}$. The mechanism that controls the size of the integration sample for each individual $i$, $|S'_i|$, is
explained in the following section.

2.5.1 Reducing the Integration Sample Error

Let \( N \) be the current data sample and \( S = \{ S_1, S_2, \ldots, S_{|N|} \} \) be the current integration sample. The integration sample variance between the parameters \( \theta_i \) and \( \theta_i^c \) is calculated using the current integration sample \( S \).

\[
\sigma^2_S(\theta_i, \theta_i^c) \approx \sum_{i \in N} \frac{1}{|S_i|} \left\{ \frac{1}{|N|^2} s^2 \left( \frac{L_{ij}(\beta'(\theta_i))}{SP_{ij}(\theta_i)} - \frac{L_{ij}(\beta'(\theta_i^c))}{SP_{ij}(\theta_i^c)} \right) \right\},
\]

where

\[
s^2 \left( \frac{L_{ij}(\beta'(\theta_1))}{SP_{ij}(\theta_1)} - \frac{L_{ij}(\beta'(\theta_2))}{SP_{ij}(\theta_2)} \right) - \frac{L_{ij}(\beta'(\theta_1))}{SP_{ij}(\theta_1)} \sum_{\nu \in S_i} \left( \frac{L_{ij}(\beta'(\theta_1))}{SP_{ij}(\theta_1)} - \frac{L_{ij}(\beta'(\theta_2))}{SP_{ij}(\theta_2)} \right)^2.
\]

Each term in the formula above is inversely proportional to the related integration sample size \( |S_i| \) of individual \( i \). In order to simplify the notation, the following can be defined:

\[
s^2_{i \text{, diff}} := s^2 \left( \frac{L_{ij}(\beta'(\theta_1))}{SP_{ij}(\theta_1)} - \frac{L_{ij}(\beta'(\theta_1^c))}{SP_{ij}(\theta_1^c)} \right).
\]

Then the integration sample variance may be rewritten as follows:

\[
\sigma^2_S(\theta_i, \theta_i^c) \approx \sum_{i \in N} \frac{1}{|S_i|} \left\{ \frac{s^2_{i \text{, diff}}}{|N|^2} \right\}.
\]

It is desirable to minimize the computational effort by allocating the minimum required integration sample size. In doing so, the constraint that the new integration sample variance be at most \( C < 1 \) times the current integration sample variance must be satisfied. This problem can be represented as follows:

\[
\min \sum_{i \in N} |S_i'|
\]

s.t. \( \sigma^2_S(\theta_i, \theta_i^c) \leq C \sigma^2_S(\theta_i, \theta_i^c) \).
Using the simplified notation for $\sigma_S^2(\theta_t, \theta_i^c)$, the problem becomes:

$$\min \sum_{i \in N} |S'_i|$$

s.t. $$\sum_{i \in N} \frac{1}{|S'_i|} \left\{ \frac{s_{i,\text{diff}}^2}{|N|^2} \right\} \leq C \sigma_S^2(\theta_t, \theta_i^c).$$

The Lagrangian function for the problem above is represented as shown below:

$$L(|S'_1|, \ldots, |S'_{|N|}|, \lambda) = \sum_{i \in N} |S'_i| + \lambda \left( C \sigma_S^2(\theta_t, \theta_i^c) - \sum_{i \in N} \frac{1}{|S'_i|} \left\{ \frac{s_{i,\text{diff}}^2}{|N|^2} \right\} \right).$$

Hence, the optimum solution for $|S'_i|$ is:

$$|S'_i| = \left[ \frac{s_{i,\text{diff}} \left( \sum_{i \in N} s_{i,\text{diff}} \right)}{C \sigma_S^2(\theta_t, \theta_i^c)|N|^2} \right].$$

### 2.6 Variance of the Estimates

The following subsections aim to provide the detailed calculations of the variances of the parameter estimates for multinomial logit models and mixed logit models, respectively. The set of observations, $\{1, 2, \ldots, I\}$, is a representative sample from an infinite population.

#### 2.6.1 Multinomial Logit Models

For a given set of observations, $\{1, 2, \ldots, I\}$, let $\hat{\beta}$ be the maximum likelihood estimator. In this case, $\hat{\beta}$ is obtained by solving the following problem:

$$\hat{\beta} := \arg \max_{\beta} LL(\beta) = \frac{1}{I} \sum_{i=1}^{I} \log(L_{ij}, (\beta)).$$

The maximum likelihood estimator $\hat{\beta}$ satisfies the following optimality condition:

$$0 = \nabla \left( \frac{1}{I} \sum_{i=1}^{I} \log L_{ij}, (\hat{\beta}) \right) = \frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij}, (\hat{\beta}). \quad (2.6.1)$$

The true parameter vector is denoted by $\beta^*$, which is the maximum likelihood estimator for the entire population. The optimality condition for $\beta^*$ implies:

$$E[\nabla \log L_{ij}(\beta^*)] = 0,$$
where the expectation is over the set of observations.

Since sampling is done from the population of individuals, it can be assumed that the terms $\nabla \log L_{ij} (\beta^*)$, which are known as scores, are also being sampled from the population. By the Central Limit Theorem,

$$\sqrt{T} \left( \frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij} (\beta^*) \right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}), \quad (2.6.2)$$

where $\mathbf{V}$ is the covariance matrix of the scores ($\nabla \log L_{ij} (\beta^*)$) in the population:

$$\mathbf{V} := E \left[ (\nabla \log L_{ij} (\beta^*)) (\nabla \log L_{ij} (\beta^*))' \right].$$

Applying the first order Taylor approximation to the right hand side of equation (2.6.1), the following holds:

$$\frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij} (\hat{\beta}) \approx \frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij} (\beta^*) + H [\hat{\beta} - \beta^*]$$

$$0 \approx \frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij} (\beta^*) + H [\hat{\beta} - \beta^*], \quad (2.6.3)$$

where

$$H := \frac{1}{I} \sum_{i=1}^{I} \nabla^2 \log L_{ij} (\beta^*).$$

Using equation (2.6.3), the following can be derived:

$$\sqrt{T} (\hat{\beta} - \beta^*) = -\sqrt{T} H^{-1} \frac{1}{I} \sum_{i=1}^{I} \nabla \log L_{ij} (\beta^*).$$

Note that using the Law of Large Numbers, $H$ converges in probability to $H$ and similarly $H^{-1}$ to $H^{-1}$, where

$$H := E \left[ \nabla^2 \log L_{ij} (\beta^*) \right].$$

Using a generalized version of Slutsky’s theorem (See Section 2.6.1.1), the following result is obtained:

$$\sqrt{T} (\hat{\beta} - \beta^*) \xrightarrow{d} \mathcal{N}(0, H^{-1} \mathbf{V} H^{-1}). \quad (2.6.4)$$
2.6.1.1 The limiting distribution of random linear functions of vectors converging to a normal distribution

Let \( \{M_n\} \) be a sequence of random matrices which converges in probability to \( M \), and \( \{h_n\} \) be a sequence of random vectors which converges in distribution to the \( N(0, \Sigma) \) distribution. Then it follows from Hall (2005) [24] page 31 that:

\[
M_n h_n \xrightarrow{d} N(0, MM').
\]

2.6.1.2 Information Identity

At the true parameter vector \( \beta^* \), the information identity states that the covariance matrix of the scores in the population, also known as Fisher Information Matrix, is equal to the Hessian matrix in the population (See Appendix B for its derivation):

\[
V = -H. \quad (2.6.5)
\]

Using the information identity, the equation (2.6.4) simplifies to:

\[
\sqrt{I}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, -H^{-1}).
\]

2.6.2 Mixed Logit Models

In mixed logit models, in order to estimate the choice probabilities \( L_{ij}(\theta) \), sampling is used for the multi-dimensional integral. In this study, not only different samples for individuals in the population are being used, but also different sample sizes. Accordingly, the simulated choice probabilities are given by:

\[
SP_{ij}(\theta) := \frac{1}{|S_i|} \sum_{\nu \in S_i} L_{ij}(\beta^*(\theta)).
\]

Let \( \theta^* \) be the true parameter vector, and \( \hat{\theta} \) be the maximum simulated likelihood estimate of the parameter vector, since the choice probabilities are simulated. Thus, \( \hat{\theta} \) is a solution to the following equation:

\[
\frac{1}{T} \sum_{i=1}^{T} \nabla \log SP_{ij}(\hat{\theta}) = 0. \quad (2.6.6)
\]
On the other hand, $\theta^*$ is a solution to the equation given below:

$$E[\nabla \log P_{ij}(\theta^*)] = 0,$$

where the expectation is taken over the population of individuals.

In order to derive the variance of the parameter vector, the function on the left hand side of equation (2.6.6), evaluated at the true parameter vector, is decomposed into the following components:

$$\frac{1}{I} \sum_{i=1}^{I} \nabla \log SP_{ij}(\theta^*) = \frac{1}{I} \sum_{i=1}^{I} \nabla \log P_{ij}(\theta^*)$$

$$+ \left( E_S \left[ \frac{1}{I} \sum_{i=1}^{I} \nabla \log SP_{ij}(\theta^*) \right] - \frac{1}{I} \sum_{i=1}^{I} \nabla \log P_{ij}(\theta^*) \right)$$

$$+ \left( \frac{1}{I} \sum_{i=1}^{I} \nabla \log SP_{ij}(\theta^*) - E_S \left[ \frac{1}{I} \sum_{i=1}^{I} \nabla \log SP_{ij}(\theta^*) \right] \right).$$

The second term on the right-hand side of the equation given above is related to the bias, whereas the third term is related to the integration sample error. The first term is the same as in equation (2.6.2). Accordingly:

$$\sqrt{I} \left( \frac{1}{I} \sum_{i=1}^{I} \nabla \log P_{ij}(\theta^*) \right) \xrightarrow{d} N(0, \mathbf{V}).$$

For the third term, the following definition is given:

$$d_i = \nabla \log SP_{ij}(\theta^*) - E_S[\nabla \log SP_{ij}(\theta^*)],$$

where $d_i$, $i \in \{1, 2, \ldots, I\}$, can be treated as draws from the population. For each individual $i$, the mean of $d_i$ is equal to zero and the variance is equal to $V_i^2/|S_i|$, where $V_i^2$ is the variance of $d_i$ when the integration sample size is just one:

$$\sqrt{I} \left( \frac{1}{I} \sum_{i=1}^{I} d_i \right) \xrightarrow{d} N(0, \frac{1}{I} \sum_{i=1}^{I} V_i^2/|S_i|).$$

Note that the integration sample error disappears as the cardinality of the set of observations $I$ increases or as integration sample sizes $|S_i|$ increase.
For the second term, which exists due to the bias, the following second order Taylor approximation is done:

\[
\nabla \log \, SP_{ij}(\theta) \approx \nabla \log \, P_{ij}(\theta) + \frac{d}{dP_{ij}(\theta)}(\nabla \log \, P_{ij}(\theta)) \left[ SP_{ij}(\theta) - P_{ij}(\theta) \right]
\]

\[
+ \frac{1}{2} \frac{d^2}{dP_{ij}(\theta)^2}(\nabla \log \, P_{ij}(\theta)) \left[ SP_{ij}(\theta) - P_{ij}(\theta) \right]^2.
\]

By calculating the expectation over the integration sample, the second term on the right-hand side disappears (unbiased choice probability estimates), which leads to:

\[
E_S[\nabla \log \, SP_{ij}(\theta)] - \nabla \log \, P_{ij}(\theta)
\]

\[
\approx \frac{1}{2} \frac{d^2}{dP_{ij}(\theta)^2}(\nabla \log \, P_{ij}(\theta)) \left[ E_S[SP_{ij}(\theta) - P_{ij}(\theta)] \right]^2
\]

\[
\approx \frac{1}{2} \frac{d^2}{dP_{ij}(\theta)^2}(\nabla \log \, P_{ij}(\theta)) \text{Var}(SP_{ij}(\theta))
\]

\[
\approx \frac{1}{2} \frac{d^2}{dP_{ij}(\theta)^2}(\nabla \log \, P_{ij}(\theta)) \frac{Q_i}{|S_i|},
\]

where \(Q_i\) is the variance of choice probability with integration sample size of one.

After summing over the set of observations, the following holds:

\[
E_S[\frac{1}{I} \sum_{i=1}^{I} \nabla \log \, SP_{ij}(\theta)] - \frac{1}{I} \sum_{i=1}^{I} \nabla \log \, P_{ij}(\theta) \approx \nabla \log \, P_{ij}(\theta) \frac{Q_i}{|S_i|}.
\]

2.6.2.1 Bias for Mixed Logit Models

The bias at parameter vector \(\theta\) in mixed logit models is calculated as follows:

\[
\text{Bias}_S^I(\theta) := SLL^S_I(\theta) - LL(\theta).
\]

If both the cardinality of set of observations \(I\) and the integration sample sizes \(|S_i|\) tend to infinity, but the integration sample sizes \(|S_i|\) tend to infinity sufficiently faster than \(I\), then the above bias term disappears. The proof of the latter is given below
following a similar direction to the proof given in Bastin (2004) [4]. Here, however, as opposed to Bastin (2004) [4], the simulation sample sizes for each individual do not have to be equal.

Let \( \delta > 0 \). From the Uniform Law of Large Numbers (ULLN), \( \exists I(\delta) \) such that \( \forall I \) s.t. \( I \geq I(\delta) \) the following holds:

\[
\sup_\theta \left| \mathbb{E}[\log P_{ij}(\theta)] - \frac{1}{I} \sum_{i=1}^{I} \log(P_{ij}(\theta)) \right| < \frac{\delta}{2} \text{ a.s.}
\]

For any given set of observations \( \{1, 2, \ldots, I\} \) and \( \delta > 0 \), again using ULLN, \( \exists S_i(\delta) \) such that \( \forall S_i \) s.t. \( |S_i| \geq S_i(\delta) \) the following holds:

\[
\sup_\theta \left| \frac{1}{I} \left( \log(P_{ij}(\theta)) - \log \left( \frac{1}{|S_i|} \sum_{\nu \in S_i} L_{ij}(\beta^\nu(\theta)) \right) \right) \right| < \frac{\delta}{2I} \text{ a.s.}
\]

Combining these inequalities using triangular inequality:

\[
\sup_\theta |SLLS_N(\theta) - LL(\theta)| \leq \sup_\theta \left| \mathbb{E}[\log P_{ij}(\theta)] - \frac{1}{I} \sum_{i=1}^{I} \log(P_{ij}(\theta)) \right| + \sum_{i=1}^{I} \sup_\theta \left| \frac{1}{I} (\log(P_{ij}(\theta)) - \log(SP_{ij}^S)) \right| \leq \delta \text{ a.s.}
\]

Let \( \{\delta_i\}_{i=1}^\infty \) be a sequence that is converging to zero. If the size of the data sample \( I' \) grows faster than \( I(\delta_i) \) and the sizes of the simulation samples \( |S_i'| \) grow faster than \( S_i(\delta_i) \) \( \forall i \in N \), then:

\[
\sup_\theta |SLLS_N(\theta) - LL(\theta)| \to 0 \text{ a.s.} \tag{2.6.7}
\]

If, on the other hand, \( I' \) grows slower than \( I(\delta_i) \), it is still possible to identify an increasing subsequence of data samples \( \{N'\} \) that grows faster than \( I(\delta_i) \). For cardinalities of set of observations \( I'(\delta_i) \) between \( I(\delta_i) \) and \( I(\delta_{i+1}) \), (2.6.7) holds if simulation sample sizes \( |S_i'| \) are equal or larger than \( S(\delta_{i+1}) \) \( \forall i \in \{1, 2, \ldots, I\} \). As a consequence, (2.6.7) holds irrespective of the speed of growth of \( I \) provided \( |S_i| \) grows sufficiently fast \( \forall i \in \{1, 2, \ldots, I\} \).
2.7 Computational Results

In this study, all three methods have been implemented and experimental results revealed that the Steihaug method performed better than the other two methods.

The software developed in order to test the algorithm of this study is given the name MLOPT, which stands for Mixed Logit Optimization Software. Actually, MLOPT is coded for solving both multinomial logit and mixed logit problems. This code is developed in C under Debian Linux. It uses a well-known linear algebra package called ATLAS [2]. The computational results are compiled with a Pentium 4 2.53Ghz PC with 1.25 GB RAM.

MLOPT is compared with AMLET [5], which is software designed to estimate multinomial and mixed logit models. This open-source software is very competitive in comparison to existing tools and includes quasi Monte-Carlo techniques based on Halton sequences developed by Bastin, Cirillo, and Toint (2003) [6]. In both AMLET and MLOPT, the Park-Miller [31] pseudorandom number generator is used.

For testing purposes, some of the datasets used for computational results in AMLET are adopted. These datasets comprise both simulated and real datasets. The simulated datasets are the same datasets generated and used in a study by Bastin, Cirillo, and Toint (2003) [6]. The attribute values are drawn from a standard univariate normal distribution. The coefficient of each independent variable is also drawn from a univariate normal distribution $N(0, 5, 1)$. The error terms are generated from the extreme value (Gumbel) distribution. There are three simulated datasets, namely “2DIM”, “5DIM”, and “10DIM”. They have 2, 5 and 10 independent normal parameters, respectively. Thus, these datasets have 4, 10 and 20 parameters to estimate, respectively. The real dataset is obtained from the six-week travel diary Mobidrive (Axhausen et al. (2002) [3]) collected in 1999 in Karlsruhe and Halle, Germany. The individuals are faced with five transportation mode alternatives, namely car driver,
car passenger, public transport, walk and bike. In [6], the authors considered the portion of the dataset related to Karlsruhe and screened the rest. The reduced dataset contains 5799 observations. There are two models to be estimated from this dataset. The first model, “Mobidrive simple”, has three independent normal parameters and 11 constant parameters, five of which are alternative-specific constants. The second model, called “Mobidrive complex”, has seven independent normal parameters and 19 constant parameters, five of which are constants.

One of the differences between MLOPT and AMLET is that AMLET uses a maximum sample size for computations, whereas MLOPT does not assume any upper bound on the sample sizes. Thus, for AMLET, four different maximum sample sizes (100, 500, 1000 and 2000) were tested. The CPU times of AMLET for all of the five problems are given in Table 2.7.1. Taking the running times from AMLET as a basis, running times selected as one of the stopping criteria for MLOPT were deliberately chosen to be much smaller. However, even if much smaller running times are used, the sample sizes in MLOPT become greater than the maximum sample sizes in AMLET. In order to compare the solutions obtained from AMLET and MLOPT, the simulated log-likelihood function values of both solutions were calculated using a big sample size (10000) with 10 different initial seeds.

MLOPT achieves bigger (better) mean simulated log-likelihood values in less time in all problem instances. The results for the simulated problems are given in Table 2.7.2.
Table 2.7.2: SLL Values for Simulated Problems

<table>
<thead>
<tr>
<th>Problem Name</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>S</td>
<td>$ (AMLET)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>2DIM MLOPT</td>
<td>-1.4996869</td>
<td>-1.4996564</td>
<td>-1.4996539</td>
<td>-1.4996539</td>
</tr>
<tr>
<td>MLOPT AMLET</td>
<td>-1.4436774</td>
<td>-1.4434793</td>
<td>-1.4434602</td>
<td>-1.4434634</td>
</tr>
<tr>
<td>5DIM MLOPT</td>
<td>-1.4051237</td>
<td>-1.4045375</td>
<td>-1.4045172</td>
<td>-1.4045003</td>
</tr>
<tr>
<td>MLOPT AMLET</td>
<td>-1.4056335</td>
<td>-1.4045712</td>
<td>-1.4045234</td>
<td>-1.4045081</td>
</tr>
</tbody>
</table>

Table 2.7.3: Relative Errors for Simulated Problems

<table>
<thead>
<tr>
<th>Problem Name</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>S</td>
<td>$ (AMLET)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>2DIM MLOPT</td>
<td>0.068192138</td>
<td>0.068211089</td>
<td>0.068212642</td>
<td>0.068212642</td>
</tr>
<tr>
<td>MLOPT AMLET</td>
<td>0.068179339</td>
<td>0.068208914</td>
<td>0.068212518</td>
<td>0.068212332</td>
</tr>
<tr>
<td>5DIM MLOPT</td>
<td>0.103009946</td>
<td>0.103126384</td>
<td>0.10312831</td>
<td>0.10312831</td>
</tr>
<tr>
<td>MLOPT AMLET</td>
<td>0.102992797</td>
<td>0.103115884</td>
<td>0.103127751</td>
<td>0.103125763</td>
</tr>
<tr>
<td>10DIM MLOPT</td>
<td>0.126947558</td>
<td>0.127311784</td>
<td>0.127324397</td>
<td>0.127334898</td>
</tr>
<tr>
<td>MLOPT AMLET</td>
<td>0.126630801</td>
<td>0.127290845</td>
<td>0.127320545</td>
<td>0.127330052</td>
</tr>
</tbody>
</table>

The relative errors for these problems obtained by AMLET and MLOPT are given in Table 2.7.3. For the maximum simulated likelihood estimate $\hat{\theta}$, the relative error is calculated as follows:

$$\text{Relative Error}(\hat{\theta}) := \frac{SLL(\hat{\theta}) - SLL(\theta_o)}{-SLL(\theta_o)},$$

where $\theta_o$ is the initial estimate of the parameter vector.

The comparison of the computation times are represented in Figures 2.7.1, 2.7.2, and 2.7.3.

The difference in the results obtained for the models on the real-life dataset, namely Mobidrive, is more dramatic. The comparison of log-likelihood values and relative errors can be seen in Table 2.7.4 and Table 2.7.5, respectively. The comparison of the computation times is represented in Figures 2.7.4 and 2.7.5.
Figure 2.7.1: Comparison of computation times in problem 2DIM.

Figure 2.7.2: Comparison of computation times in problem 5DIM.

Figure 2.7.3: Comparison of computation times in problem 10DIM.
### Table 2.7.4: SLL Values for Mobidrive Problems

| Problem Name | | | | | |
|--------------|--------|--------|--------|--------|
|              | $|S|$ (AMLET) |
|              | 100    | 500    | 1000   | 2000   |
| MD\_simple   | MLOPT  | -1.1647184 | -1.1647185 | -1.1647131 | -1.1647129 |
|              | AMLET  | -1.1648790 | -1.1647247 | -1.1647139 | -1.1647131 |
| MD\_complex  | MLOPT  | -1.1045942 | -1.1045646 | -1.1045696 | -1.1045620 |
|              | AMLET  | -1.1046048 | -1.1045722 | -1.1045698 | -1.1045639 |

### Table 2.7.5: Relative Errors for Mobidrive Problems

| Problem Name | | | | | |
|--------------|--------|--------|--------|--------|
|              | $|S|$ (AMLET) |
|              | 100    | 500    | 1000   | 2000   |
| MD\_simple   | MLOPT  | 0.174471793 | 0.174471722 | 0.17447555 | 0.174475691 |
|              | AMLET  | 0.174357963 | 0.174467328 | 0.174474983 | 0.17447555 |
| MD\_complex  | MLOPT  | 0.217482756 | 0.217503726 | 0.217500184 | 0.217505568 |
|              | AMLET  | 0.217475247 | 0.217498342 | 0.217500042 | 0.217504222 |

### Figure 2.7.4: Comparison of computation times in problem MDs.

59
Figure 2.7.5: Comparison of computation times in problem MDc.
CHAPTER III

APPLICATION TO AUTOMOTIVE INDUSTRY

A new trust-region based algorithm was introduced for solving mixed logit models in Chapter 2. The computational results for this algorithm were very encouraging for both simulated and real-life data. The next step in this study is to develop a mixed logit demand model for an application. This mixed logit demand model will be benchmarked against another well-known discrete choice demand model, namely a multinomial logit model.

The focus of this application is on the automotive industry, specifically the new car market. First, a multinomial logit model is developed. Its results are presented and its deficiencies in modeling demand patterns (e.g., unrealistic substitution patterns) are discussed. Next, two mixed logit models are developed. The mixed logit models developed in this study are able to predict market shares and explain the behaviors of individuals under different scenarios as explained in detail in this chapter. More specifically, the responsiveness of individuals to changes in a given scenario is considered. For instance, in a case where an additional rebate for an alternative (e.g., a rebate for an American car) is introduced, individuals having different car preferences (taste variations) will react differently to this rebate, meaning that an individual who likes American cars may be strongly affected by this change in scenario whereas an individual preferring European cars may not respond at all. Such behaviors that are incorporated in the model are called substitution patterns. In such settings in which different substitution behaviors of individuals are to be incorporated, there is an important advantage to using mixed logit models. This advantage will be demonstrated by presenting the abovementioned models with substitution patterns. An objective of
this chapter is to compare the extent to which the mixed logit and multinomial logit
models are able to predict the market shares and the responsiveness of the individuals
to changes in the financing options.

3.1 Data Sources

The discrete choice models in this study characterize new vehicle purchasers using a
utility function that relates the purchasers’ choices to the characteristics of the vehicle
(both physical and financial) and the characteristics of the purchaser himself. In order
to pursue this modeling approach, a number of sources were used to gather data.
These sources included sales transaction data, vehicle characteristic data, quality and
reliability data, and census data. These data sources are described in this section.

A dataset of new vehicle transactions was obtained from General Motors (GM),
which includes vehicle features and sales transactions. Vehicle features include the
model, make, model year, etc. Some of the transaction features are sales amount
(sales price), purchase decision (purchased or leased), rebate, trade allowance, etc.
See Appendix D for a list of the data fields available from the transaction data.

There are additional vehicle attributes (e.g., fuel consumption, horsepower, inte-
rior passenger volume) that potentially affect purchase decisions. These attributes
were not available in the transaction data, but were obtained from an auxiliary
database of vehicle attributes and features provided by General Motors. See Ap-
pendix D for a listing of the data available from this source.

The quality (or perceived quality) of the vehicle is one of the factors that affect the
purchase decision. In order to include this factor in the demand models, the initial
quality survey and vehicle dependability study from J. D. Power and Associates were
used. Both of these attributes are measured for each vehicle model and different
model year, in problems per hundred vehicles (PPH).

One of the assumptions in the discrete choice models is that the error terms $\varepsilon_{ij}$
in the utility function in (1.4.2) have zero mean. In almost all cases, the attributes used in modeling the utility function do not suffice to model the utility such that the error terms have zero mean. Therefore, alternative specific indicator variables are generally used to model that part of the utility that cannot be incorporated in the model by the attributes. However, if these indicators are included in the utility function for the vehicle models, the physical vehicle attributes and quality metrics cannot be used in the demand model directly, because this creates linear dependence between the set of vehicle model indicators and the vehicle attributes, meaning that each vehicle attribute vector can be written as a linear combination of the set of vehicle model indicators. In order to include these attributes in the demand model without creating dependence, the vehicle attributes need to have variation within the same vehicle model. This variation exists among different trim levels of the same vehicle model. By using the vehicle attributes of the specific trim level and model year purchased by the individuals, this dependence is avoided in this study. Similar variation exists when the physical vehicle attributes are interacted with the attributes of the individuals (e.g. Price/Income Level) and the dependence is avoided.

In order to merge information from all these data sources, the different make and model naming conventions had to be harmonized. Both the vehicle attributes and sales transaction data sources are indexed in terms of vehicle make and model names. However, the model names used in these data sources are different. For instance, for BMW 325I, "325I" is used as the model name in the sales transaction data source, whereas "BMW 325I" is used in the vehicle attributes data source. Also, the quality of the data needed to be thoroughly examined. The outliers and the transaction records containing errors were removed. In joint work with the Global Market and Industry Analysis (GMIA) group in General Motors, criteria were identified and applied to screen out measurements on attributes that were inconsistent and/or unreliable. For instance, the transactions with unusually high (a new Cavalier sale with a sale price
of $99,999) or low (a new Mini Cooper sale with a sale price of $3000) sale prices were removed from the analysis.

Prior to developing discrete choice models the entire transaction record was further examined to verify its consistency for modeling purposes; that is, does the total vehicle price (including fees, insurance, etc.) agree with the total amount the customer paid (including down payment, trade-in equity, etc.). Records that were inconsistent were not used in the discrete choice modeling. This additional screening step eliminated a small percentage of the total number of records, and approximately one million records were retained for the modeling. The modeling data consisted of sales data covering a multi-year time span. The sales transaction attributes are different for cash, finance and lease transactions, and they also vary depending on whether or not there is a trade-in vehicle (See Section 3.2.3 for details).

3.2 Constructing the Demand Model

Mixed logit models are used for predicting demand or market share as well as predicting the effect of changes in the attribute values on demand under different scenarios. In this study, the market of interest is the new car market. Thus, the conditional probability of new car purchases is estimated, which excludes the important used car market and the people who delay their purchase or decide not to buy. Nevertheless, this approach is well accepted and frequently applied in the literature (Train (2007) [41]). Alternatively, several scholars in the literature have proposed using an outside good alternative (Berry, Levinsohn, and Pakes (1995) [12]), but in this case, the lack of data necessary to assign characteristics to the artificial alternative prevents the use of such an approach.

Customer-level data is one of the most important elements of any demand model. The transaction data for different products in the automotive industry obtained from General Motors not only contains new vehicle sales transactions for GM vehicles but
also for their competitors. Using this data, a discrete choice model, more specifically a mixed logit model, has been developed. A specific segment of the new car market called Low Segment was selected for the analysis. However, the application of the mixed logit model of this study for other segments or combination of segments is similar. As mentioned in Chapter 1, in mixed logit models it is assumed that the consumers' preferences over the attributes of the demand model, represented by the coefficients $\beta$, vary with a known distribution. For that reason, the parameters $\theta$ of the distribution of these preferences $\beta$ have to be estimated.

The new car market segment used in this study, namely the Low Segment, contains 15 different vehicle models. These models are given in Table 3.2.1 with their observed market share values. The cleaned data for the Low Segment contains 69406 records. The choice sets constructed for the individuals depend on the region of the consumer and the time of the purchase. The construction of these choice sets is explained in Section 3.2.1. The sizes of the choice sets are around 100 each. For each consumer, all of the alternatives in his choice set need to be specified. Hence, the input file constructed for the discrete choice model has almost seven million rows, meaning that it is a very large dataset.

Most of the discrete choice analyses for the automotive industry in the literature have concentrated on the consumer choices between different make-model vehicles. See Train and Winston (2007) [41], Berry, Levinsohn, and Pakes (1995) [12], and Sudhir (2001) [39]. However, transaction types such as cash/finance purchase or lease option have different revenue implications for the automakers. For this purpose, transaction type is included as one of the alternative dimensions in the demand models of this study. Each transaction in the sales transaction data corresponds to two parts; a consumer, who will be denoted by $i$, and the actual choice, $j_i$, made by that individual $i$. The choice selected (or the alternative purchased) by that individual will be represented in terms of the following characteristics:
Table 3.2.1: Low Segment - Market Share Data

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.95</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>9.05</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.72</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.92</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>13.78</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.28</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.45</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.76</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.02</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.32</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.10</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.25</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.62</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.01</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.76</td>
</tr>
</tbody>
</table>

- Make (e.g., Toyota, Chevrolet)
- Car model (e.g., Corolla, Cavalier)
- Transaction type (e.g., Lease 36 months, Finance 60 months, Cash)

Hence, different values of the triplet \{ Make, Model, Transaction Type \} correspond to different alternatives. One example of such an alternative is:

Alternative \( j = \{ \text{Chevrolet, Cobalt, Finance 60 months} \}. \)

The definition of an alternative might be extended further by, for instance, including the trim level of the vehicle purchased as an additional alternative dimension. The trade-off is the extra computational burden resulting from the larger choice sets for the individuals.

Accordingly, the same vehicle model purchased with different transaction types will be treated as different alternatives in the model. For instance, a Chevrolet Cobalt vehicle purchased with cash will be treated as a different alternative than a Chevrolet Cobalt vehicle financed for 60 months. Each transaction in the sales data corresponds
to a purchase made using one of the transaction types. The initial analysis on the sales transactions showed that most of the sales were accomplished using five different transaction types. These are:

- Cash purchase
- Financing 60 months
- Financing 72 months
- Leasing 36 months
- Leasing 48 months

3.2.1 Choice Set Construction

One of the assumptions of discrete choice models is that each individual faces a finite set of alternatives and each individual chooses only one from this set. These alternatives and their attributes are visible to everybody (both decision makers and the demand model analyst). Decision makers do not have to have the same choice sets. However, their alternative sets must be finite and they must be visible.

The demand model analysts generally have some kind of transaction dataset detailing the choices made by each individual in this dataset. Thus, what demand model analysts generally observe is the choices that individuals make and some of the attributes of these choices. However, in order to model the discrete choice system, the other alternatives available to each individual as well as the attributes of these alternatives need to be known. There is a need for constructing a choice set for each individual.

Different choice sets for different regions and time periods (month of the year) have been constructed for this purpose. This is a reasonable assumption since the automakers' price/incentive programs change from month to month and from region to region. For each region and month of the year, it is assumed that the choice set
consists of the alternatives actually selected by the individuals (as obtained from the sales transaction data) during that time period and within that region. As regards the attributes of these alternatives, the average of each attribute value in the sales transaction data is used within that region and time period. This will be a crude approximation, since, for instance, every individual gets a different financing rate for a new vehicle depending on their credit score.

Even though the alternative sets are constructed using average attribute values, the real attributes of the alternative actually selected by the individuals are supplied from the sales transaction dataset. For all other alternatives that are not selected, the average attribute values remain as their attribute values.

### 3.2.2 Utility Function Attributes

For the choice sets explained in Subsection 3.2.1, five different attributes are employed. These attributes are Quality, Price of Accessories and Options (PAO), Residual Value Percentage of Vehicle, Operating Cost and Net Present Value (NPV). The detailed explanations of these attributes are given in the following subsections.

#### 3.2.2.1 Quality

For each vehicle model and for different model years, the Initial Quality Study (IQS) and the Vehicle Dependability Study (VDS) data were obtained from J.D. Power and Associates. Initial Quality Study (IQS) is a study used for quality benchmarking in the auto industry and it focuses on problems that consumers report with their new vehicles during the first three months of ownership. The Vehicle Dependability Study (VDS), on the other hand, measures long-term quality after a few years of ownership. This study surveys new vehicle owners regarding problems experienced after four to five years of ownership. The study is not only based on vehicle defects, but also on the interplay of factors such as wear and tear, mileage and price. The importance of this quality criterion in the choices of consumers comes from the fact that the study
provides insight into the reliability and dependability of brands and models as they approach the end of a typical warranty period.

The IQS and VDS are measured in problems per hundred vehicles (PPH). It is preferable to incorporate both of these criteria in the demand model. In order to relate them as a single attribute to the vehicle in question, a way of combining the IQS and the VDS is to take their average:

\[ \text{Quality} := \frac{\text{IQS} + \text{VDS}}{2}. \]

As such, this new quality definition becomes a selection criterion for the vehicle and represents an attribute affecting the utility of the vehicle for the consumers in the demand models.

3.2.2.2 Price of Accessories and Options (PAO)

For each transaction, there are two price related pieces of data available. The first one is the (manufacturer’s) suggested retail price (MSRP), which is the price the manufacturer recommends that the retailer sell that vehicle. The second price data is the base (manufacturer’s) suggested retail price (BaseMSRP), which is the recommended retail price of a product in its simplest form, i.e. without any accessories or options. The difference between these two prices will provide, for each transaction, the value in price of the accessories or options available on that particular vehicle. Hence, the values of the options purchased by the individuals can be calculated as:

\[ \text{PAO} := \text{MSRP} - \text{BaseMSRP}. \]

3.2.2.3 Residual Value Percentage

Residual Value describes the future value of the vehicle in terms of a percentage of depreciation of its initial value. For instance, if a vehicle is sold at a list price of $20,000 today, after a usage of 24 months its value will be reduced, due to depreciation, to $13,000. Hence, in this example, the Residual Value Percentage is the
percent ratio of the vehicle’s value after 2 years of depreciation (i.e. $13,000) to the
vehicle’s initial value (i.e. $20,000), which is 65%. The Residual Value Percentages
exist for different periods of usage such as 24 months or 36 months.

3.2.2.4 Operating Cost

The operating cost of a vehicle is, as the name implies, how much it costs to operate
a car (also called variable costs or out of pocket expenses), such as fuel, oil, tire wear,
maintenance, which increases with vehicle use and is given in terms of operating
cost/mile. The operating cost that is used as a vehicle attribute in this study consists
of the dollar value of the fuel consumption of the car for each mile. Hence, Operating
Cost will be calculated by:

\[
\text{OPCOST} := \frac{\text{Fuel Price} (\$/\text{gallon})}{\text{MPG} (\text{miles/gallon})},
\]

where MPG is the miles per gallon fuel consumption of the related vehicle, and Fuel
Price is the average price of gas in the region that the individual lives.

3.2.2.5 Net Present Value

Depending on the type of the purchase, the sales transaction record contains inform-
ation on different attributes. Cash transactions have the Sale Amount and the
Rebate fields. For finance and lease transactions, the necessary fields are Down Pay-
ment, Trade Equity, Monthly Payments, and Purchase Term. On top of these, lease
transactions will have a Residual Amount. These fields define the cash flow used for
the purchase. Using an internal rate of return, the cash flow can be discounted into
today’s monetary value so that different transactions will have a comparable measure
for price. Net present values of the transactions are calculated differently for different
transaction types and the equations of NPV for each transaction type are given in
Section 3.2.3. In the demand model, three NPV related attributes are used. These
are denoted as \( \text{NPV}_{\text{cash}} \), \( \text{NPV}_{\text{finance}} \), and \( \text{NPV}_{\text{lease}} \).
3.2.3 Net Present Value Calculations

The type of the purchase (cash, finance, lease) as described in the previous subsection determines the payment method for each consumer. However, in order to be able to compare these different purchase methods, they are converted to a relative value at a common point in time. Thus, their net present values are calculated.

For cash transactions, the payments are done up front in cash and/or trade-equity. For finance and lease transactions, the consumers generally pay an upfront down payment in cash and/or trade-equity, followed by monthly payments of equal sizes. For lease transactions, in particular, at the end of the lease term, the vehicle may either be returned to the dealer or purchased by the consumer for the residual value written in the contract. The payments made in cash and/or trade-equity are treated as cash flow and the present value of these cash flows can be evaluated for each transaction. Based on the borrower’s opportunity cost of money, the present value of the cash flow will be the net present value (NPV) of the sales transaction (Dasgupta, Siddarth, and Silva-Risso (2007) [22]). Borrowers’ opportunity cost is defined as the cost incurred by the loss of money-making opportunities (by market interest rate) when the individuals make a new car purchase decision. Hence, NPV represents the financial cost of the purchase not only in terms of the retail price but also the opportunity cost of the underlying cash flow of the investment. For instance, for finance transactions, the monthly payments are discounted by a common rate, which is a predetermined market rate supplied by GM. Hence, the net present value of the investment related to each alternative is calculated. Depending on the transaction type of the purchase, the related net present value calculation differs.

Some of the purchases involve trade-in vehicles, in which case the consumers trade-in their old vehicles, and purchase or lease a new vehicle. For these cases, the NPV calculations include an additional term representing the value of that trade-in. In order to calculate the value of the trade-in, there are a few terms that need to be
considered.

The consumer may owe money on the trade-in vehicle that he is trading. The amount that the consumer owes on the trade-in vehicle is called the Payoff. The actual cash value of the trade-in vehicle is represented by Trade ACV. The dealer may buy the trade-in vehicle for the Trade ACV. However, the dealer may buy the old car over or under the Trade ACV. Depending on the situation, the dealer may give more than the Trade ACV as an incentive for the consumer. This difference is represented by the Trade Over Value. Note that if this attribute is negative, then it means that the dealer gets the trade-in vehicle at a lower value than its actual cash value. The Trade Amount is the amount given for the trade-in vehicle, which is the sum of the Trade ACV and the Trade Over Value. Using these terms, the equity of the trade-in vehicle, denoted by True ACV, can be calculated as follows:

\[
\text{True ACV} := \text{Trade ACV} - \text{Payoff}.
\]

For cash transactions, the net present value is calculated as follows:

\[
\text{NPV}_{\text{cash}} := \text{Cash Down} + \text{True ACV}.
\]

For finance and lease transactions, the consumers pay monthly payments and these payments must be converted into today’s monetary value. For this purpose, each monthly payment needs to be discounted by the interest rate. This interest rate is taken to be equal to the prime interest rate plus one percent. Suppose a consumer pays monthly payments of \$A for \(t\) months. Using an interest (discount) rate of \(r\), the net present value of these monthly payments will be:

\[
\text{PV}(A, r, t) := A \left( \frac{1 - \left(1 + \frac{r}{t}\right)^{-t}}{r} \right).
\]

For finance transactions, the net present value formula is:

\[
\text{NPV}_{\text{finance}} := \text{Cash Down} + \text{True ACV} + \text{PV(Monthly Payment, } r, \text{ Term)}.
\]

72
Specifically for lease transactions, at the end of the contract the car is returned to the dealer or purchased by the consumer by the Residual Amount. Accordingly, in order to incorporate this Residual Amount at the end of the contract, it needs to be discounted. Hence, the net present value formula for lease transactions is:

\[
\text{NPV}_{\text{lease}} := \text{Cash Down} + \text{True ACV} + \text{PV} (\text{Monthly Payment}, r, \text{Term}) \\
+ \text{Residual Amount} \times (1 + r)^{-t}.
\]

### 3.3 Choice Probabilities & Market Shares

In this section, choice probabilities for different discrete choice models will be reviewed. The main focus will be on calculating the market share or market volumes using the choice probabilities of the individuals.

#### 3.3.1 Multinomial Logit Model

For multinomial logit models, the \(i^{th}\) individual’s choice probability for alternative \(j\) as a function of the relative importance vector \(\beta\) is given by:

\[
L_{ij}(\beta) := \frac{\exp(\beta'x_{ij})}{\sum_{i \in C_i} \exp(\beta'x_{il})}.
\]

The market shares can be calculated by using the individuals’ choice probabilities. Letting \(s_j(\beta)\) denote the predicted market share for alternative \(j\):

\[
s_j(\beta) := \frac{1}{I} \sum_{i=1}^{I} L_{ij}(\beta). \quad (3.3.1)
\]

#### 3.3.2 Mixed Logit Model

For mixed logit models, the \(i^{th}\) individual’s choice probability function for alternative \(j\) as a function of the parameter vector \(\theta\) is given by:

\[
P_{ij}(\theta) := \int L_{ij}(\beta) f(\beta|\theta) \, d\beta.
\]
Generally, the multidimensional integrals in the choice probabilities shown above cannot be calculated in closed form, but can be approximated by their Monte Carlo estimates. Using the integration sample $S$, these actual choice probabilities $P_{ij}(\theta)$ are approximated by the simulated choice probabilities:

$$SP_{ij}^{S}(\theta) := \frac{1}{|S_i|} \sum_{\nu \in S_i} L_{ij}(\beta^\nu(\theta)).$$

The term $L_{ij}(\beta^\nu(\theta))$ in the above equation is the choice probability function for the multinomial logit case, when $\beta = \beta^\nu(\theta)$.

The market shares can be estimated by using the individuals’ simulated choice probabilities. Letting $s^S_j(\theta)$ denote the predicted market share for alternative $j$:

$$s^S_j(\theta) := \frac{1}{I} \sum_{i=1}^I SP_{ij}^{S}(\theta). \quad (3.3.2)$$

### 3.3.3 Market Size

It will be assumed that the size of the new vehicle market is equal to a constant $M$. An immediate drawback of this assumption is evident under the scenario where the prices of all new vehicles increase while the prices for used vehicles remain constant. In this case, ideally, a vehicle choice model would predict a shrinking new vehicle market size. Even though this assumption of constant market size seems to be restrictive, the market size is generally well predicted and stable. Thus, it can be an input to the demand model framework.

Depending on the type of the discrete choice demand model used in the analysis, the market volume will be calculated differently. According to this assumption, the predicted market volume, $q_j$, for alternative $j$ will be:

- Multinomial Logit

$$q_j := Ms_j(\beta),$$

where $s_j(\beta)$ is defined in equation (3.3.1).
• Mixed Logit

\[ q_j := M s^2_j(\theta), \]

where \( s^2_j(\theta) \) is defined in equation (3.3.2).

In the equations above, the constant market size is \( M \), and the size of the transaction data used for testing is \( I \).

3.4 Demand Model Results

In this section, the comparison of mixed logit models with multinomial logit models will be presented. With each of these discrete choice models, demand in the automotive industry will be predicted using the datasets explained in Section 3.1. These models are constructed with similar sets of attributes, as explained in Section 3.2.2. In order to analyze the substitution patterns of these models, a scenario of an additional rebate on the Cavalier cash purchases is constructed and applied. The substitution patterns as well as penetration rates are discussed in this section. To finalize, a complex mixed logit model is developed in order to demonstrate the superior aspects of mixed logit models in predicting real-life demand for automotive industry.

3.4.1 Multinomial Logit Model

In order to construct a multinomial logit demand model, several attributes are first selected to be included in the utility function. These attributes are Operating Cost ($/mile), Quality (PPH), Residual Percentage, Price of Accessories / Options (PAO) as well as Net Present Value. Three different Net Present Value attribute categories are used depending on the type of the sales transaction (\( NPV_{\text{cash}}, NPV_{\text{finance}}, NPV_{\text{lease}} \)). On top of these, indicator variables for 14 out of 15 vehicle models are adopted. The indicator variable for one of the vehicle models is set to zero. These indicator variables are used to model the average utility not measured by the above mentioned attributes. It is a general practice to set one of the indicator variables to
zero, since in discrete choice models only differences in the utility matter (not the magnitudes of the utilities). An initial analysis on the transaction data showed that the Quality of the vehicle model purchased and the income level of the individual have positive correlation. A similar correlation exists between the Residual Percent of the vehicle model purchased and the education level of the individual. Education level of the individuals are measured in terms of number of years at school, whereas the income level is measured in terms of natural logarithm of the yearly household income. In order to capture these correlations, two new variables are created by multiplying Residual Percent and Quality variables by Education Level and Income Level variables respectively.

The results of the parameter estimates are given in Table 3.4.1. Note that all of the parameters are statistically significant, since their standard deviations are very small. There are a few remarks worth mentioning regarding the parameter estimates. To begin with, the signs of the parameter estimates are an important aspect when assessing the quality of the estimates, since the parameter estimates are actually an indicator of a taste preference. As such, one would expect the parameter estimate for the price attribute (NPV) to be negative, since clearly an increase in price (NPV) would surely decrease the utility function value of an individual. On the other hand, the parameter estimate of an attribute like the Residual Percentage is expected to be positive, which follows from the fact that the higher the Residual Percentage of the car, the better the investment return associated with buying the car (because the re-sale value would be proportionally higher), and therefore the higher the utility function value of the individual. Based on these remarks, it can be concluded that the parameter estimates obtained in the results presented have the expected signs. Looking at Table 3.4.1, for instance, the parameter estimates for the NPV attributes are negative whereas the parameter estimate for the PAO attribute is positive.
<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Parameter Estimate</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Cost</td>
<td>-1.9659</td>
<td>0.0262</td>
</tr>
<tr>
<td>Quality × Log Income</td>
<td>-0.0178</td>
<td>0.0003</td>
</tr>
<tr>
<td>Options</td>
<td>0.1757</td>
<td>0.0051</td>
</tr>
<tr>
<td>Residual Percent × Education</td>
<td>0.0095</td>
<td>0.0002</td>
</tr>
<tr>
<td>NPV_{cash}</td>
<td>-0.1432</td>
<td>0.0027</td>
</tr>
<tr>
<td>NPV_{finance}</td>
<td>-0.1324</td>
<td>0.0023</td>
</tr>
<tr>
<td>NPV_{lease}</td>
<td>-0.2086</td>
<td>0.0026</td>
</tr>
<tr>
<td>Beetle</td>
<td>0.5356</td>
<td>0.0987</td>
</tr>
<tr>
<td>Cavalier</td>
<td>5.0834</td>
<td>0.0869</td>
</tr>
<tr>
<td>Civic</td>
<td>-2.9697</td>
<td>0.0835</td>
</tr>
<tr>
<td>Cobalt</td>
<td>4.4925</td>
<td>0.0889</td>
</tr>
<tr>
<td>Corolla</td>
<td>-5.0885</td>
<td>0.1089</td>
</tr>
<tr>
<td>Elantra</td>
<td>-0.9073</td>
<td>0.0293</td>
</tr>
<tr>
<td>Jetta</td>
<td>1.8649</td>
<td>0.0772</td>
</tr>
<tr>
<td>Matrix</td>
<td>-2.3468</td>
<td>0.0682</td>
</tr>
<tr>
<td>Mazda3</td>
<td>-0.0802</td>
<td>0.0690</td>
</tr>
<tr>
<td>Mini Cooper</td>
<td>-0.7202</td>
<td>0.1018</td>
</tr>
<tr>
<td>Sentra</td>
<td>0.9518</td>
<td>0.0379</td>
</tr>
<tr>
<td>Vibe</td>
<td>1.6033</td>
<td>0.0522</td>
</tr>
<tr>
<td>tC</td>
<td>-1.1178</td>
<td>0.0784</td>
</tr>
<tr>
<td>xB</td>
<td>2.3328</td>
<td>0.1070</td>
</tr>
</tbody>
</table>
Table 3.4.2: Multinomial Logit Model - Market Share Results

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Observed Market Share</th>
<th>Estimated Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.95</td>
<td>2.97</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>9.05</td>
<td>8.95</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.72</td>
<td>17.19</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.92</td>
<td>4.60</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>13.78</td>
<td>14.80</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.28</td>
<td>6.91</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.45</td>
<td>10.27</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.76</td>
<td>7.54</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.02</td>
<td>3.23</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.32</td>
<td>5.56</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.10</td>
<td>3.15</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.25</td>
<td>6.50</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.62</td>
<td>2.68</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.01</td>
<td>2.90</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.76</td>
<td>2.74</td>
</tr>
</tbody>
</table>

In order to estimate the multinomial logit demand model, 75% of the sales transaction data is randomly selected. This constitutes the model dataset. Hence, the model dataset is used to estimate the demand model parameters, i.e., the relative importance vector $\beta$. Once the multinomial logit demand model has been estimated, it has to be tested. The remaining 25% of the sales transaction data is used as the test data for evaluating the quality of the demand model. The estimated multinomial logit model predicts market shares on the test data. These predictions are then compared to the observed market shares of the test dataset. The results of the market share predictions using the multinomial logit demand model are given in Table 3.4.2.

The estimated market shares and the observed market shares match closely. Thus, the multinomial logit demand model developed accurately predicts the market shares. However, another important aspect of the demand models is how individuals react to different scenarios such as additional rebates on specific vehicle models. These behaviors are called substitution patterns.

In order to observe the substitution patterns in this multinomial logit model, an
Table 3.4.3: Multinomial Logit Model - Market Share Results Scenario1

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Market Share Base</th>
<th>Market Share Scenario</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.97</td>
<td>2.95</td>
<td>-0.54</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>8.95</td>
<td>9.46</td>
<td>5.71</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.19</td>
<td>17.10</td>
<td>-0.53</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.60</td>
<td>4.57</td>
<td>-0.59</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>14.80</td>
<td>14.72</td>
<td>-0.52</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>6.91</td>
<td>6.87</td>
<td>-0.57</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.27</td>
<td>10.21</td>
<td>-0.57</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.54</td>
<td>7.50</td>
<td>-0.54</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.23</td>
<td>3.21</td>
<td>-0.54</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.56</td>
<td>5.53</td>
<td>-0.56</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.15</td>
<td>3.14</td>
<td>-0.52</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.50</td>
<td>6.46</td>
<td>-0.57</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.68</td>
<td>2.67</td>
<td>-0.56</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>2.90</td>
<td>2.89</td>
<td>-0.56</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.74</td>
<td>2.73</td>
<td>-0.56</td>
</tr>
</tbody>
</table>

additional $1000 rebate on cash purchases of the Cavalier model is applied. Using the parameter estimates obtained from the base case, the market shares are re-estimated taking into account this additional rebate of $1000. The comparison of the market shares for the base case and the market shares for the scenario in which this additional rebate is applied is given in Table 3.4.3. The rebate has an unrealistic effect on the market shares. All of the alternatives except the Cavalier cash purchase, including the Cavalier finance purchases, exhibit an equal percent loss in market share, meaning that all the other alternatives lose approximately 0.56% of their original market shares. The market share of total Cavalier sales increases by 5.71%. This increase is low according to the estimates of the analysts in GM, indicating that in this multinomial logit demand model, the parameter estimates for the NPV attributes could be too small in magnitude.

Another important aspect of the demand model to be considered under the rebated cash purchase scenario is the penetration rate changes for the Cavalier model. Penetration rates for a vehicle model are the percent sales in different transaction
Table 3.4.4: Multinomial Logit Model - Penetration Rates for Cavalier

<table>
<thead>
<tr>
<th></th>
<th>Base Case</th>
<th>After Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Share (Cash)</td>
<td>3.70</td>
<td>4.24</td>
</tr>
<tr>
<td>Market Share (Finance)</td>
<td>5.25</td>
<td>5.22</td>
</tr>
<tr>
<td>Cash %</td>
<td>41.34</td>
<td>44.82</td>
</tr>
<tr>
<td>Finance %</td>
<td>58.66</td>
<td>55.18</td>
</tr>
</tbody>
</table>

types with respect to its overall sales. For instance, considering a car model that has a market share of 8%, which can be subdivided as 2% cash sales and 6% finance sales, the penetration rates will be 25% and 75% for the cash and finance sales, respectively. In such a setting, if there is a change in the percentage of the cash sales category due to a rebate on cash purchase, then the penetration rates will change accordingly, shifting towards a higher percentage for the cash sales and a lower percentage for the finance sales.

Changes in the penetration rates predicted by the demand model are very important. The reason for this is that manufacturers aim to increase their profitabilities by changing, either increasing or decreasing, the specific penetration rates of their vehicles, and the demand model should be able to accurately estimate the changes in the penetration rates under different scenarios. For instance, GM may want to decrease the lease penetration rate for Cobalt vehicles. In such a case, the demand model should be capable of capturing the realistic substitution patterns that would occur, for instance, under a cash incentive scenario. For this multinomial logit model, the penetration rates for the Cavalier model in the base case and under the rebate on cash purchase scenario are given in Table 3.4.4.

Looking at Table 3.4.3, it can be seen that under the rebate on cash purchase scenario, the increase in the market share of the Cavalier model is compensated by an equal percent loss of market share for all other alternatives, where percent loss is defined as the percent decrease with respect to the initial market shares. Therefore,
the conclusion is that for this multinomial logit model, the reason for the changes in the penetration rates is solely due to the market share increase in the Cavalier cash purchases. However, in reality what is expected is that the percent loss in the market share of some alternatives, such as the Cavalier finance purchase alternative, be relatively higher than the percent loss in the market shares of other alternatives.

The substitution pattern and the changes in the penetration rates observed in these results seem unrealistic and not acceptable for the purpose of this study even though the base case market share predictions are very accurate. Also, according to the estimates of the analysts in GM, the percent increase in the market share of the Cavalier vehicle model should range between $15 - 20\%$, whereas this multinomial logit model predicts a market share increase of $5.71\%$.

### 3.4.2 Mixed Logit Models

Two different mixed logit models with several attributes are estimated using MLOPT. Some of the attributes used to model the utility functions are interacted with consumers’ characteristics, examples of which are Quality $\times$ Log Income and Residual Percentage $\times$ Education Level. The other attributes are Operating Cost, $NPV_{\text{cash}}$, $NPV_{\text{finance}}$, $NPV_{\text{lease}}$, and PAO. Operating cost depends on the fuel prices in the region where the consumer lives, whereas Net Present Values depend on the financing rate that the consumer gets on the finance and lease options. On top of these attributes, indicator variables are used in the utility function for different vehicle models. Variables other than the vehicle model specific indicators, on the other hand, will vary for each vehicle model depending on the consumer.

#### 3.4.2.1 Model # 1

The first model developed in this section demonstrates the construction of a substitution pattern in mixed logit models. For this purpose, a simple mixed logit model is presented. See Section 1.6.3.1 for a detailed discussion about substitution patterns
in Mixed Logit models. This model uses all of the attributes used in the multinomial logit example given in Section 3.4.1. For the multinomial logit model presented in that example, all of the parameters are constant. In mixed logit model #1, all the parameters, with one exception only, will be kept constant. Hence, only one attribute will be selected to have a non-constant parameter. The selected parameter is modeled as normally distributed with unknown mean and standard deviation. The purpose of this simple mixed logit model is to demonstrate the construction of a substitution pattern using random parameters $\beta$. Suppose it is believed that the substitution pattern between different vehicles depends only on the Operating Cost of the vehicles. For this purpose, the selected attribute with the non-constant relative importance parameter is the Operating Cost. Hence, the aim is to construct a model in which alternatives that have similar operating costs will compete against each other more aggressively than against the alternatives that are quite different in operating cost values.

The results of the parameter estimates for the base case are given in Table 3.4.5. Note that all of the parameters are statistically significant, since their standard deviations shown in the third column are very small. For the operating cost attribute, since its parameter is selected to be non-constant, instead of a single constant value, a set of distribution parameters are being estimated. Since the selected parameter is chosen to be normally distributed, its unknown mean and standard deviation are estimated. These estimates have their respective standard deviations, too, as can be seen in Table 3.4.5. The results shown in this table indicate that estimates for the mean and standard deviation for the parameter distribution of the Operating Cost attribute are not only significant but also big in magnitude. This indicates that there is considerable taste variation for the corresponding attribute, i.e. Operating Cost, in the population.

As discussed in Chapter 2, the stopping criterion used for MLOPT depends on
### Table 3.4.5: Mixed Logit Model # 1 - Results

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Parameter Estimate</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Cost (Mean)</td>
<td>-2.0126</td>
<td>0.0360</td>
</tr>
<tr>
<td>Operating Cost (Std)</td>
<td>3.7450</td>
<td>0.06295</td>
</tr>
<tr>
<td>Quality × Log Income</td>
<td>-0.0283</td>
<td>0.0003</td>
</tr>
<tr>
<td>PAO</td>
<td>0.0886</td>
<td>0.0037</td>
</tr>
<tr>
<td>Residual Percent × Education</td>
<td>0.0076</td>
<td>0.0002</td>
</tr>
<tr>
<td>$NPV_{\text{cash}}$</td>
<td>-0.1544</td>
<td>0.0022</td>
</tr>
<tr>
<td>$NPV_{\text{finance}}$</td>
<td>-0.1408</td>
<td>0.0018</td>
</tr>
<tr>
<td>$NPV_{\text{lease}}$</td>
<td>-0.2222</td>
<td>0.0022</td>
</tr>
<tr>
<td>Beetle</td>
<td>3.1969</td>
<td>0.1136</td>
</tr>
<tr>
<td>Cavalier</td>
<td>8.1450</td>
<td>0.0999</td>
</tr>
<tr>
<td>Civic</td>
<td>-4.2618</td>
<td>0.0706</td>
</tr>
<tr>
<td>Cobalt</td>
<td>7.6199</td>
<td>0.1077</td>
</tr>
<tr>
<td>Corolla</td>
<td>-7.9467</td>
<td>0.1006</td>
</tr>
<tr>
<td>Elantra</td>
<td>-0.7259</td>
<td>0.0311</td>
</tr>
<tr>
<td>Jetta</td>
<td>3.6477</td>
<td>0.0884</td>
</tr>
<tr>
<td>Matrix</td>
<td>-1.4551</td>
<td>0.0672</td>
</tr>
<tr>
<td>Mazda3</td>
<td>2.1900</td>
<td>0.0796</td>
</tr>
<tr>
<td>Mini Cooper</td>
<td>2.5772</td>
<td>0.1146</td>
</tr>
<tr>
<td>Sentra</td>
<td>2.3848</td>
<td>0.0511</td>
</tr>
<tr>
<td>Vibe</td>
<td>3.4894</td>
<td>0.0700</td>
</tr>
<tr>
<td>tC</td>
<td>-2.0353</td>
<td>0.1083</td>
</tr>
<tr>
<td>xB</td>
<td>6.1727</td>
<td>0.1296</td>
</tr>
</tbody>
</table>
the gradient norm square error $\sigma_{||g^*(\theta)||^2}$. Using this stopping criterion, the outcome of the computations obtained with MLOPT for the model presented in this section is the parameter vector given in Table 3.4.5. When the algorithm stopped, the average integration sample size was around 3500, the gradient norm square error was $9 \times 10^{-5}$, and the relative error was 0.1226.

Using the sales transaction data, two datasets are created. First, 75% of the sales transactions are randomly selected for fitting the mixed logit model. This constitutes the model dataset. Then, the remaining 25% of the sales transactions are used as the test data. The estimated mixed logit model predicts market shares for the test data. The market shares estimated by the mixed logit model are then compared to the observed market shares in the test dataset. This comparison is presented in Table 3.4.6.

The results presented in Table 3.4.6 show that the mixed logit demand model estimated for the Low Segment is quite successful in predicting market shares. Based on this conclusion, the next step in the analysis is to test the behavior of this model.

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Observed Market Share</th>
<th>Estimated Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.95</td>
<td>2.92</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>9.05</td>
<td>9.03</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.72</td>
<td>17.58</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.92</td>
<td>4.99</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>13.78</td>
<td>14.21</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.28</td>
<td>7.33</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.45</td>
<td>10.56</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.76</td>
<td>7.21</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.02</td>
<td>3.02</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.32</td>
<td>5.38</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.10</td>
<td>3.06</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.25</td>
<td>6.14</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.01</td>
<td>3.19</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.62</td>
<td>2.62</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.76</td>
<td>2.76</td>
</tr>
</tbody>
</table>
under a different scenario. Before that, the substitution pattern constructed in the mixed logit model should be analyzed.

First, the utility function for this model is formulated. The following terminology is defined for the Operating Cost attribute in this formulation:

- The Operating Cost attribute, denoted by \( z_{ij}^{\text{OPCOST}} \),

- The component of the relative importance vector related to Operating Cost, denoted by \( \gamma_i^{\text{OPCOST}} \),

- The mean of the component of the relative importance vector related to Operating Cost (assumed to be normally distributed), denoted by \( \mu^{\text{OPCOST}} \),

- The standard deviation of the component of the relative importance vector related to Operating Cost (assumed to be normally distributed), denoted by \( \sigma^{\text{OPCOST}} \).

The other attributes are represented by the vector \( x_{ij} \) and their relative importance vector is given by \( \beta \). Accordingly, the utility function takes the following form:

\[
U_{ij}(x_{ij}, z_{ij}^{\text{OPCOST}}, \beta, \gamma_i^{\text{OPCOST}}) := \beta' x_{ij} + \gamma_i^{\text{OPCOST}} z_{ij}^{\text{OPCOST}} + \varepsilon_{ij}.
\]

Under this setting, consider three alternatives, namely \( j \) (for Cavalier), \( k \) (for Cobalt) and \( l \) (Corolla). Cavalier and Cobalt have similar and high Operating Cost values, whereas Corolla has a considerably lower Operating Cost with respect to the other two alternatives. Operating Cost value is calculated by Fuel Price divided by MPG of the vehicles. MPG values for Cavalier and Cobalt are around 25, whereas for Corolla it is around 32. Having such three alternatives, suppose that there is an additional rebate for Cavalier. In this case, it is normally expected that as Cavalier increases its market share, it should steal relatively more share from Cobalt than from Corolla. The reason for such an expectation is that since Cobalt is a more similar
competitor in terms of its Operating Cost, it will be affected more heavily from a rebate for Cavalier.

This expected behavior can be justified by comparing the cross-elasticity of Cobalt to the cross-elasticity of Corolla when there is an additional rebate for Cavalier. Using equation (1.6.13), the cross-elasticity of choice probability for Cobalt when NPV of Cavalier is increased by one percent is calculated by:

\[ E_{i,\text{Cobalt},\text{NPV}_{\text{Cavalier}}} = -\beta^{\text{NPV}_{\text{NPV}_{\text{Cavalier}}}} P_{i,\text{Cavalier}}(\theta) \left( 1 + \text{Cov} \left( \frac{L_{i,\text{Cavalier}}(\beta)}{P_i,\text{Cavalier}(\theta)}, \frac{L_{i,\text{Cobalt}}(\beta)}{P_i,\text{Cobalt}(\theta)} \right) \right). \]

Likewise, the cross-elasticity for Corolla is calculated as follows:

\[ E_{i,\text{Corolla},\text{NPV}_{\text{Cavalier}}} = -\beta^{\text{NPV}_{\text{NPV}_{\text{Cavalier}}}} P_{i,\text{Cavalier}}(\theta) \left( 1 + \text{Cov} \left( \frac{L_{i,\text{Cavalier}}(\beta)}{P_i,\text{Cavalier}(\theta)}, \frac{L_{i,\text{Corolla}}(\beta)}{P_i,\text{Corolla}(\theta)} \right) \right). \]

Since Cavalier and Cobalt have similar operating costs and Corolla has a lower operating cost:

\[ \text{Cov} \left( \frac{L_{i,\text{Cavalier}}(\beta)}{P_i,\text{Cavalier}(\theta)}, \frac{L_{i,\text{Cobalt}}(\beta)}{P_i,\text{Cobalt}(\theta)} \right) > \text{Cov} \left( \frac{L_{i,\text{Cavalier}}(\beta)}{P_i,\text{Cavalier}(\theta)}, \frac{L_{i,\text{Corolla}}(\beta)}{P_i,\text{Corolla}(\theta)} \right). \]

Therefore \( |E_{i,\text{Cobalt},\text{NPV}_{\text{Cavalier}}}| > |E_{i,\text{Corolla},\text{NPV}_{\text{Cavalier}}}|. \) Since the magnitude of the cross-elasticity of Cobalt is higher than that of Corolla, Cobalt is more sensitive to the price change of Cavalier.

In order to test whether the substitution pattern constructed by the above model works as intended, the same scenario used in the multinomial logit example in Section 3.4.1 is adopted. Hence, an additional cash rebate of $1000 is applied to the cash purchases for Cavalier vehicle models. The resulting market shares and the penetration rates are analyzed to assess the demand prediction quality of mixed logit models and to observe the structure of the substitution pattern constructed. The effect of these scenarios is compared with the estimates of the analysts in GM.

The comparison of market shares under the base case and the scenario is demonstrated in Table 3.4.7. The most striking difference between this comparison and
the one presented in Table 3.4.3 for the multinomial logit example is that the % change in the market shares are different for each vehicle model. Not only are the % changes in the market shares for each vehicle model different, but they also differ from each other in a logical and expected way. For instance, the results obtained show that the Cavalier cash purchase alternative steals more market share from the Cavalier finance purchase alternative than from all the other remaining alternative vehicles. Moreover, it is observed that the vehicle models having Operating Cost values similar to the Cavalier model exhibit a higher percent loss in market share than the vehicle models that are less similar in terms of Operating Cost. Vehicle models such as Cobalt, Elantra, Focus, Matrix, Mazda3, Mini Cooper, Sentra, Vibe and xB have Operating Cost values similar to the Cavalier model. Hence, their market share loss is more than the other vehicles. Models like Corolla, Civic and tC have much lower Operating Cost values than the Cavalier model, and hence they are the least affected by this additional rebate. This behavior is anticipated by the construction of the model in this example. Accordingly, it can be concluded that the mixed logit model developed in this example is a good model for predicting demand having a substitution pattern based on operating cost, although the percent changes in market shares are low in magnitude when compared with those estimated by analysts within GM.

Even though the substitution pattern that is constructed by this mixed logit demand model seems good, it nevertheless has a problem. The % increase in the market share of the Cavalier vehicles (5.92%) is low compared to the 15-20% increase in market share estimated by GM analysts. This is mainly due to the low magnitudes of the NPV related parameters in the estimated model. However, the signs of all of the estimated parameters are as expected.

The penetration rates in the base case and in the case of the Cavalier cash purchase rebate scenario are shown in Table 3.4.8. The penetration rates change in favor of
Table 3.4.7: Mixed Logit Model # 1 - Market Share Results Scenario

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Market Share Base</th>
<th>Market Share Scenario</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.92</td>
<td>2.91</td>
<td>-0.45</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>9.03</td>
<td>9.57</td>
<td>5.92</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.58</td>
<td>17.50</td>
<td>-0.48</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.99</td>
<td>4.96</td>
<td>-0.65</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>14.21</td>
<td>14.15</td>
<td>-0.46</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.33</td>
<td>7.28</td>
<td>-0.76</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.56</td>
<td>10.48</td>
<td>-0.72</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.21</td>
<td>7.18</td>
<td>-0.45</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.02</td>
<td>3.00</td>
<td>-0.75</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.38</td>
<td>5.34</td>
<td>-0.71</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.06</td>
<td>3.04</td>
<td>-0.71</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.14</td>
<td>6.09</td>
<td>-0.77</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.19</td>
<td>3.18</td>
<td>-0.20</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.62</td>
<td>2.60</td>
<td>-0.76</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.76</td>
<td>2.74</td>
<td>-0.75</td>
</tr>
</tbody>
</table>

Table 3.4.8: Mixed Logit Model # 1- Penetration Rates for Cavalier

<table>
<thead>
<tr>
<th></th>
<th>Base Case</th>
<th>After Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Share (Cash)</td>
<td>3.70</td>
<td>4.28</td>
</tr>
<tr>
<td>Market Share (Finance)</td>
<td>5.34</td>
<td>5.29</td>
</tr>
<tr>
<td>Cash %</td>
<td>40.92</td>
<td>44.70</td>
</tr>
<tr>
<td>Finance %</td>
<td>59.08</td>
<td>55.30</td>
</tr>
</tbody>
</table>

a shift in the cash purchases of the Cavalier model, as expected. Compared to the penetration rate changes of the multinomial logit model example given in Table 3.4.3, where the increase in the market share of the Cavalier model is compensated by an equal percent loss of market share for all other alternatives, the penetration rate changes in the simple mixed logit model of this example are slightly higher.

3.4.2.2 Model # 2

In mixed logit model # 1 constructed in Section 3.4.2.1, there is only one attribute that has a non-constant parameter and all the other attributes have constant parameters. The main aim of that simple example was to demonstrate the construction
of a substitution pattern between alternatives by specifying the random parameters in the mixed logit model. In this subsection, a more complex model with a more complicated substitution pattern is constructed.

The same set of attributes are used in this model as in the multinomial logit example and the simple mixed logit example. However, this time, the parameters for the Operating Cost attribute and the three NPV attributes (NPV\textsubscript{cash}, NPV\textsubscript{finance}, NPV\textsubscript{lease}) are assumed to be random with normal distributions. The results of the parameter estimates are given in Table 3.4.9.
All of the constant parameters are statistically significant. Also, all of the random attributes have significant means and standard deviations. Note that the standard deviations of the random parameters are not only statistically significant but are also quite big in magnitude, suggesting that there is considerable taste variation for the corresponding attributes in the population. Also the signs of the parameter estimates are meaningful. Another important observation for the parameter estimates is that the means of the NPV related attributes are larger in magnitude when compared to their corresponding estimates in the multinomial logit model example and the simple mixed logit model example (Model \# 1). Hence, this implies that rebates will have a greater impact on market shares in this complex mixed logit model since the changes caused by the rebates affect the utility function to a greater extent. All this occurs due to the larger average magnitude of the parameters in this mixed logit model. In the previous examples the NPV related attributes had small magnitude estimates for the means and as a result, the percent market share changes seemed small.

As discussed in Chapter 2, the stopping criterion used for MLOPT depends on the gradient norm square error \( \sigma_{||g^k(\theta)||^2} \). Using this stopping criterion, the outcome of the computations obtained with MLOPT for the model presented in this subsection is the parameter vector given in Table 3.4.9. When the algorithm stopped, the average integration sample size was around 3500, the gradient norm square error was \( 1.8 \times 10^{-4} \), and the relative error was 0.1476.

Using the sales transaction data, two datasets are created. First, 75% of the sales transactions are randomly selected for fitting the mixed logit model. This constitutes the model dataset. Then, the remaining 25% of the sales transactions are used as the test data. The estimated mixed logit model predicts market shares on the test data, and these estimates are then compared to the observed market shares in the test dataset. This comparison is presented in Table 3.4.10.

The results presented in Table 3.4.10 show that the mixed logit demand model
Table 3.4.10: Mixed Logit Model # 2 - Market Share Results

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Observed Market Share</th>
<th>Estimated Market Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.95</td>
<td>2.63</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>9.05</td>
<td>8.73</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.72</td>
<td>17.80</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.92</td>
<td>4.81</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>13.78</td>
<td>14.09</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.28</td>
<td>7.88</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.45</td>
<td>10.48</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>7.76</td>
<td>6.56</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>3.02</td>
<td>2.96</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.32</td>
<td>5.31</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.10</td>
<td>3.86</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.25</td>
<td>6.26</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.62</td>
<td>2.57</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.01</td>
<td>3.23</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.76</td>
<td>2.84</td>
</tr>
</tbody>
</table>

is quite successful in predicting market shares for the Low Segment. Based on this conclusion, the next step in the analysis is to test the behavior of this model under a scenario. Before that, however, the substitution pattern constructed in this mixed logit model will be discussed.

Since the parameter for the Operating Cost attribute (OpCost) and the three NPV attributes (NPV\textsubscript{cash}, NPV\textsubscript{finance}, NPV\textsubscript{lease}) are assumed to be random with normal distributions, the substitution pattern that will be observed by this example is expected to depend on these attributes. This means that there will be correlations between the alternatives that have similar NPV and Operating Cost attributes. The three NPV related attributes are defined as the product of the NPV attribute with the appropriate indicator variable for the type of the purchase (cash, finance, lease). According to this definition the NPV related attributes are written in the following form:

- NPV\textsubscript{cash} := NPV × Indicator Variable\textsubscript{cash},
\[ \text{NPV}_{\text{finance}} := \text{NPV} \times \text{Indicator Variable}_{\text{finance}}; \]

\[ \text{NPV}_{\text{lease}} := \text{NPV} \times \text{Indicator Variable}_{\text{lease}}. \]

Thus, the interactions between the alternatives will not merely be based on the NPV attribute but also on the purchase type of the alternative. In order to test the constructed substitution pattern, the same scenario used in the former examples is applied to this mixed logit model. In this scenario, an additional cash rebate of $1000 is applied to the cash purchases for the Cavalier vehicle model. The resulting market share changes are given in Table 3.4.11. Similar to the simple mixed logit example, the percent changes in the market shares are different for each vehicle model. The market share of the overall Cavalier vehicle model increases by 22.9%. This number is more realistic based on the estimates of the analysts in GM (15%-20%). This percent increase in the market share can be attributed to the larger magnitudes of the NPV related attribute parameters.

The Cavalier vehicle models captured more market share from Elantra, Sentra, xB, Cobalt and Focus than from the other models. These vehicle models have Operating Cost attributes and NPV attributes similar to Cavalier, which explains the reason why their market share losses are greater than the rest. Mini Cooper, Vibe, Matrix and Mazda3 have similar Operating Cost attribute values, and accordingly, in the simple mixed logit example (Model \# 1) they lost more market share in comparison to the other vehicle models. However, in this model their market share losses are moderate. This is because of the fact that they have higher NPV values than Cavalier even though they have similar Operating Cost values. The market share losses for Civic and Corolla are also moderate even though these vehicles do not have Operating Cost values similar to Cavalier. This is mainly due to the fact that their NPV values are closer to the NPV values of Cavalier. Hence, in this example, it is possible to observe the interaction of the Operating Cost and NPV related attributes.
Table 3.4.11: Mixed Logit Model # 2 - Market Share Results Scenario

<table>
<thead>
<tr>
<th>Vehicle Make</th>
<th>Vehicle Model</th>
<th>Market Share Base</th>
<th>Market Share Scenario</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen</td>
<td>Beetle</td>
<td>2.63</td>
<td>2.63</td>
<td>-0.06</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cavalier</td>
<td>8.73</td>
<td>10.73</td>
<td>22.93</td>
</tr>
<tr>
<td>Honda</td>
<td>Civic</td>
<td>17.80</td>
<td>17.56</td>
<td>-1.24</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>Cobalt</td>
<td>4.81</td>
<td>4.65</td>
<td>-3.34</td>
</tr>
<tr>
<td>Toyota</td>
<td>Corolla</td>
<td>14.09</td>
<td>13.83</td>
<td>-1.81</td>
</tr>
<tr>
<td>Hyundai</td>
<td>Elantra</td>
<td>7.88</td>
<td>7.36</td>
<td>-6.60</td>
</tr>
<tr>
<td>Ford</td>
<td>Focus</td>
<td>10.48</td>
<td>10.18</td>
<td>-2.88</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>Jetta</td>
<td>6.56</td>
<td>6.54</td>
<td>-0.22</td>
</tr>
<tr>
<td>Toyota</td>
<td>Matrix</td>
<td>2.96</td>
<td>2.93</td>
<td>-0.93</td>
</tr>
<tr>
<td>Mazda</td>
<td>Mazda3</td>
<td>5.31</td>
<td>5.27</td>
<td>-0.71</td>
</tr>
<tr>
<td>Mini</td>
<td>Mini Cooper</td>
<td>3.86</td>
<td>3.86</td>
<td>0.00</td>
</tr>
<tr>
<td>Nissan</td>
<td>Sentra</td>
<td>6.26</td>
<td>5.94</td>
<td>-5.17</td>
</tr>
<tr>
<td>Scion</td>
<td>tC</td>
<td>3.23</td>
<td>3.21</td>
<td>-0.47</td>
</tr>
<tr>
<td>Pontiac</td>
<td>Vibe</td>
<td>2.57</td>
<td>2.55</td>
<td>-0.62</td>
</tr>
<tr>
<td>Scion</td>
<td>xB</td>
<td>2.84</td>
<td>2.73</td>
<td>-3.87</td>
</tr>
</tbody>
</table>

Table 3.4.12: Mixed Logit Model # 2 - Penetration Rates for Cavalier

<table>
<thead>
<tr>
<th></th>
<th>Base Case</th>
<th>After Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Share (Cash)</td>
<td>2.98</td>
<td>5.07</td>
</tr>
<tr>
<td>Market Share (Finance)</td>
<td>5.75</td>
<td>5.66</td>
</tr>
<tr>
<td>Cash %</td>
<td>34.17</td>
<td>47.24</td>
</tr>
<tr>
<td>Finance %</td>
<td>65.83</td>
<td>52.76</td>
</tr>
</tbody>
</table>

The penetration rates in the base case and in the case of the Cavalier cash purchase rebate scenario are shown in Table 3.4.12. The penetration rates change in favor of a shift in the cash purchases of the Cavalier model, as expected. However, in comparison to the penetration rate changes in the simple mixed logit model example shown in Table 3.4.8, it is clearly seen that the changes obtained in the more complex mixed logit model of this example are more dramatic.
CHAPTER IV

CONCLUSION AND FUTURE RESEARCH

The algorithm proposed in this study is a new trust-region based algorithm that solves mixed logit problems by maximizing the simulated log-likelihood function in (2.1.4), which furthermore employs statistical testing for evaluating the quality of the generated solutions. The algorithm adaptively changes the sizes of both the data sample ($N$) and the integration (simulation) sample ($S$). The statistical testing for the quality of the generated solutions enables the algorithm to stop when the approximate solution is satisfactory.

Computational results for both simulated and real-life data are very encouraging. An important application to the auto industry is presented. Mixed logit demand models with realistic substitution patterns are constructed and estimated with the data obtained from the auto industry. A discussion of how to create realistic substitution patterns by specifying the random parameters of the utility function is presented. The demand model results and a comparison between the different discrete choice models is given. A discussion of the advantages and shortcomings of the multinomial logit and mixed logit models of this study is also provided.

The mixed logit models developed in this study are novel from many different aspects as compared to the demand models available in the literature. To begin with, the structure of the model is different due to the way the alternatives are defined. Depending on the transaction type, the sales of each vehicle model has a corresponding revenue implication for the manufacturers, which can be any one of the cash, finance (with different contract lengths) and lease (with different contract lengths) options. Therefore, the mixed logit models developed for this study have transaction
type dependent alternatives. Accordingly, due to this nature of the alternatives, the constructed demand model is able to predict the market shares or penetration rates for the different transaction types for the vehicles. These demand models and the market share predictions obtained by the models constitute an important tool for the manufacturer, which not only explains the profitability of the vehicle sales for the manufacturer's product portfolio, but also develops a methodology to increase the profitability (overall or for individual vehicles in terms of financing options) by correctly estimating market shares and penetration rates for different what-if scenarios.

Another novel aspect introduced by this study is to consider the Price of Accessories and Options (PAO) as an attribute in the utility function. The PAO attribute is an important factor affecting demand and its exclusion may lead to a deficiency of the discrete choice models called endogeneity. In a sales transaction data, the prices (NPV) of the vehicles that individuals choose to buy differ because of several reasons. The two most important reasons behind these differences is that every individual gets different financing rates for the same vehicle and every individual has different options and accessories installed in the purchased vehicle. Naturally, these extras always make the prices of the vehicles higher. Even though these loaded vehicles are priced higher than the base model vehicles, they are desirable. Thus, not considering these extra accessories and options as an attribute in the utility function will cause biased estimates for the price parameters, meaning that they may turn out to be negative but small in magnitude or even positive, which is not as expected. This study has successfully included PAO related attributes in the demand models and removed the related endogeneity inherent in these demand models.

In the specification of the random parameters, the distribution type used for the mixed logit models developed in this study was iid normal. However, there are also other distribution types that are found to be very useful. For instance, the relative importance parameter for the price (NPV) attribute is always expected to be
non-positive. Hence, including different distribution types, such as log-normal, will improve the model specification. The log-normally distributed parameter may be a promising extension for the specification of the NPV parameters.

This study has been applied over a particular segment, namely the Low Segment of the automotive industry. As a future extension, this mixed logit model framework should also be applied to different segments or even to a combination of several segments.
APPENDIX A

MULTINOMIAL LOGIT CHOICE PROBABILITY

There are different derivations of the choice probabilities in multinomial logit models. The derivation given here is very similar to that in Train (2002) [40]. Another derivation can be found in Ben-Akiva and Lerman (1985) [10]. In multinomial logit models, the relative importance vector $\beta$ is assumed to be constant, and the error terms are Gumbel distributed. The cumulative density function of the Gumbel distribution, with the location parameter $\eta = 0$ and the scale parameter $\mu = 1$ is given by:

$$F(t) := \exp(-\exp(-t)),$$

whereas the density function is given by:

$$f(t) = \exp(-\exp(-t)) \exp(-t).$$

Using the utility maximization assumption, the choice probability is defined by:

$$L_{ij}(\beta) := \Pr(\beta'x_{ij} + \varepsilon_{ij} > \beta'x_{il} + \varepsilon_{il}, \ \forall l \in C_i \ & l \neq j)$$

$$= \Pr(\beta'x_{ij} - \beta'x_{il} + \varepsilon_{ij} > \varepsilon_{il}, \ \forall l \in C_i \ & l \neq j).$$

Conditioning on $\varepsilon_{ij}$ and using the fact that the error terms $\varepsilon_{il}$ are distributed independent Gumbel, the above equation is the product of cumulative Gumbel functions:

$$L_{ij}(\beta|\varepsilon_{ij}) = \prod_{l \in C_i \ & l \neq j} \exp(-\exp(-(\beta'x_{ij} - \beta'x_{il} + \varepsilon_{ij}))).$$

The choice probability $L_{ij}(\beta)$ can be evaluated by taking the integral over the
distribution of $\varepsilon_{ij}$:

$$L_{ij}(\beta) = \int L_{ij}(\beta|\varepsilon_{ij}) \exp(-\exp(-\varepsilon_{ij})) \exp(-\varepsilon_{ij}) d\varepsilon_{ij}$$

$$= \int \left[ \prod_{l \neq j} \exp(-\exp(-\beta' x_{ij} - \beta' x_{il} + \varepsilon_{ij})) \right] \exp(-\exp(-\varepsilon_{ij})) \exp(-\varepsilon_{ij}) d\varepsilon_{ij}.$$

For $l = j$, the following holds:

$$\exp(-\exp(-\beta' x_{ij} - \beta' x_{il} + \varepsilon_{ij})) = \exp(-\exp(-\varepsilon_{ij})).$$

Using this equality,

$$\left[ \prod_{l \neq j} \exp(-\exp(-\beta' x_{ij} - \beta' x_{il} + \varepsilon_{ij})) \right] \exp(-\exp(-\varepsilon_{ij})) = \prod_{l \in C_i} \exp(-\exp(-\beta' x_{ij} - \beta' x_{il} + \varepsilon_{ij})).$$

Hence, the choice probability can be rewritten as:

$$L_{ij}(\beta) = \int \left[ \prod_{l \in C_i} \exp(-\exp(-\beta' x_{ij} - \beta' x_{il} + \varepsilon_{ij})) \right] \exp(-\varepsilon_{ij}) d\varepsilon_{ij}$$

$$= \int \exp \left( -\exp(-\varepsilon_{ij}) \sum_{l \in C_i} \exp(-\beta' x_{ij} - \beta' x_{il}) \right) \exp(-\varepsilon_{ij}) d\varepsilon_{ij}.$$

Using change of variable by $t = -\exp(-\varepsilon_{ij})$ and $\exp(-\varepsilon_{ij}) d\varepsilon_{ij} = dt$, the choice probability becomes:

$$L_{ij}(\beta) = \int_0^\infty \exp(-t \sum_{l \in C_i} \exp(-\beta' x_{ij} - \beta' x_{il})) dt$$

$$= \frac{\exp(-t \sum_{l \in C_i} \exp(-\beta' x_{ij} - \beta' x_{il}))}{\sum_{l \in C_i} \exp(-\beta' x_{ij} - \beta' x_{il})} \bigg|_{t=0}^\infty$$

$$= 1$$

$$= \frac{\exp(\beta' x_{ij})}{\sum_{l \in C_i} \exp(\beta' x_{il})}.$$
APPENDIX B

INFORMATION IDENTITY

The score function is the derivative of the logarithm of the likelihood function with respect to the parameter $\beta$. The score function for the individual $i$ for selecting the alternative $j$ is denoted by:

$$\nabla \log L_{ij}(\beta).$$

At the true parameter vector $\beta^*$, the information identity states that the covariance matrix of the scores in the population, also known as the Fisher Information Matrix, is equal to the Hessian of the scores with respect to the parameter $\beta$ in the population.

The set of individuals, $\{1, 2, \ldots, I\}$, in this study is finite. However, the derivations for an infinite set of individuals is straightforward (See Train (2002) [40]). In this setting, the summation of the choice probabilities for each individual $i$ over their alternatives in their choice sets adds up to $I$. Hence:

$$1 = \frac{1}{I} \sum_{i=1}^{I} \sum_{j \in C_i} L_{ij}(\beta).$$

The derivative of the above equation with respect to the relative importance vector is:

$$0 = \frac{1}{I} \sum_{i=1}^{I} \sum_{j \in C_i} \nabla L_{ij}(\beta).$$

Using the identity:

$$\nabla L_{ij}(\beta) = \nabla \log L_{ij}(\beta) \ L_{ij}(\beta),$$

the above equation can be rewritten as:

$$0 = \frac{1}{I} \sum_{i=1}^{I} \sum_{j \in C_i} \nabla \log L_{ij}(\beta) \ L_{ij}(\beta).$$
Taking the derivative of the above equation with respect to the relative importance vector $\beta$:

$$0 = \frac{1}{T} \sum_{i=1}^{T} \sum_{j \in C_i} L_{ij}(\beta) \nabla^2 \log L_{ij}(\beta) + \nabla \log L_{ij}(\beta) \nabla L_{ij}(\beta)'$$

$$= \frac{1}{T} \sum_{i=1}^{T} \sum_{j \in C_i} (\nabla^2 \log L_{ij}(\beta) + \nabla \log L_{ij}(\beta) \nabla \log L_{ij}(\beta)') L_{ij}(\beta).$$

Hence,

$$-\frac{1}{T} \sum_{i=1}^{T} \sum_{j \in C_i} \nabla^2 \log L_{ij}(\beta) L_{ij}(\beta) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j \in C_i} \nabla \log L_{ij}(\beta) \nabla \log L_{ij}(\beta)' L_{ij}(\beta).$$

In the above equation, the left-hand side is the negative of the average Hessian in the set of individuals, whereas the right-hand side is the variance of the scores in the set of individuals:

$$-H = V.$$
APPENDIX C

GRADIENT AND HESSIAN CALCULATIONS

Both for the multinomial logit and mixed logit models, gradient and Hessian calculations of the (simulated) log-likelihood function are very important. These calculations take up about 99% of the execution time in MLOPT.

C.1 Multinomial Logit

The log-likelihood function is given in equation (1.6.3). The gradient of the log-likelihood function, denoted by \( g(\beta) \in \mathbb{R}^K \), is calculated as follows:

\[
g(\beta) := \nabla LL(\beta) = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{L_{ij_i}(\beta)} \nabla L_{ij_i}(\beta).
\] (C.1.1)

For additional notation, let \( X_i \) be a \( K \times |C_i| \) matrix having the attribute vectors of the \( i^{th} \) individual’s choice set as its columns:

\[
\overline{X}_i := \begin{bmatrix} \bar{x}_{i1} & \bar{x}_{i2} & \ldots & \bar{x}_{i|C_i|} \end{bmatrix}.
\]

For individual \( i \), the calculated choice probabilities for each alternative in the choice set \( C_i \) forms the choice probability vector \( \mathcal{T}_i(\beta) \):

\[
\mathcal{T}_i(\beta) := \begin{bmatrix} L_{i1}(\beta) \\ L_{i2}(\beta) \\ \vdots \\ L_{i|C_i|}(\beta) \end{bmatrix}.
\]

The choice probability vector \( \mathcal{T}_i(\beta) \) and its components are functions of the parameter vector \( \beta \). The gradient of the choice probabilities with respect to \( \beta \) is given
by:

\[
\nabla L_{ij}(\beta) = L_{ij}(\beta) \left[ x_{ij} - \sum_{l \in C_i} L_{il}(\beta) \bar{x}_l \right] \\
= L_{ij}(\beta) \bar{x}_i (\bar{e}_{ji} - \bar{T}_i(\beta)),
\]

where \(\bar{e}_{ji}\) is the \(j_i\)th column of the identity matrix.

Using equations (C.1.1) and (C.1.2), gradient vector \(g(\beta)\) can be rewritten as:

\[
g(\beta) := \frac{1}{T} \sum_{i=1}^{T} \bar{x}_i (\bar{e}_{ji} - \bar{T}_i(\beta))
\]

\[
= \frac{1}{T} \sum_{i=1}^{T} \bar{x}_i \text{res}_{ij}(\beta),
\]

where \(\text{res}_{ij}(\beta) = \bar{e}_{ji} - \bar{T}_i(\beta)\).

For Hessian calculations, additional notation is needed. Let \(\tilde{L}_i(\beta)\) be a diagonal matrix. At the diagonal, it stores the choice probabilities of the \(i\)th individual’s choice set:

\[
\tilde{L}_i(\beta) := \begin{bmatrix}
L_{i1}(\beta) & 0 & \ldots & 0 \\
0 & L_{i2}(\beta) & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & L_{i|C_i|}
\end{bmatrix}.
\]

The Hessian of the log-likelihood function can be calculated as follows:

\[
H(\beta) := \nabla^2 LL(\beta)
\]

\[
= \frac{1}{T} \sum_{i=1}^{T} \left\{ -\frac{1}{(L_{ij}(\beta))^2} \nabla L_{ij}(\beta) \nabla L_{ij}(\beta)' + \frac{1}{L_{ij}(\beta)} \nabla^2 L_{ij}(\beta) \right\}
\]

\[
= \frac{1}{T} \sum_{i=1}^{T} \left\{ \frac{1}{L_{ij}(\beta)} \nabla^2 L_{ij}(\beta) - \bar{x}_i \text{res}_{ij}(\beta) (\bar{x}_i \text{res}_{ij}(\beta))' \right\}. \tag{C.1.5}
\]

In the above Hessian calculations, the Hessian of the choice probabilities are
needed and they are calculated as:

\[
\nabla^2 L_{ij}(\beta) = \nabla \left[ L_{ij}(\beta) \left( \bar{x}_{ij} - \sum_{l \in C_i} L_{il}(\beta) \bar{x}_l \right) \right] \\
= L_{ij}(\beta) \left[ \bar{x}_{ij} - \sum_{l \in C_i} L_{il}(\beta) \bar{x}_l \right] (\mathbf{X}_i \text{ res}_{ij_i}(\beta))' \\
- L_{ij}(\beta) \sum_{l \in C_i} L_{il}(\beta) \bar{x}_i (\bar{e}_i - \bar{L}_i(\beta)) \bar{x}_l' \\
= L_{ij}(\beta) \left\{ \mathbf{X}_i \text{ res}_{ij_i}(\beta) (\mathbf{X}_i \text{ res}_{ij_i}(\beta))' - \sum_{l \in C_i} L_{il}(\beta) \bar{x}_i (\bar{e}_i - \bar{L}_i(\beta)) \bar{x}_l' \right\} \\
= L_{ij}(\beta) \left\{ \mathbf{X}_i \text{ res}_{ij_i}(\beta) (\mathbf{X}_i \text{ res}_{ij_i}(\beta))' + \bar{x}_i \bar{L}_i(\beta) (\mathbf{X}_i \bar{L}_i(\beta))' \\
- \mathbf{X}_i \bar{L}_i(\beta) \mathbf{X}_i \right\}. \\
\tag{C.1.6}
\]

Substituting (C.1.6) into (C.1.5), the Hessian matrix calculation can be simplified to:

\[
H(\beta) = \frac{1}{I} \sum_{i=1}^{I} \left\{ \mathbf{X}_i \bar{L}_i(\beta) (\mathbf{X}_i \bar{L}_i(\beta))' - \mathbf{X}_i \bar{L}_i(\beta) \mathbf{X}_i \right\}. \\
\tag{C.1.7}
\]

### C.2 Mixed Logit

In this subsection, the gradient and the Hessian of the simulated log-likelihood function \( SLL^S_N(\theta) \) will be calculated using the data sample \( N \) and the integration sample \( S \). It is assumed that the set of individuals is represented by \( \{1, 2, \ldots, I\} \).

The simulated log-likelihood function is a function of the parameter vector \( \theta \):

\[
SLL^S_N(\theta) := \frac{1}{|N|} \sum_{i \in N} \log (SP^S_{ij_i}(\theta)). \\
\tag{C.2.1}
\]

The gradient of the simulated log-likelihood function can be calculated by:

\[
g^S_N(\theta) := \nabla SLL^S_N(\theta) \\
= \frac{1}{|N|} \sum_{i \in N} \frac{1}{SP^S_{ij_i}(\theta)} \nabla SP^S_{ij_i}(\theta), \\
\tag{C.2.2}
\]

where the derivatives of the simulated choice probabilities and the conditional choice
probabilities are required:

\[
\nabla S P_{ij}^S (\mathbf{\theta}) = \frac{1}{|S_i|} \sum_{\nu \in S_i} \nabla L_{ij} (\mathbf{\beta}^\nu (\mathbf{\theta}))
\]

and

\[
\nabla P_{ij} (\mathbf{\beta}^\nu (\mathbf{\theta})) = \nabla \mathbf{\beta}^\nu (\mathbf{\theta}) \nabla L_{ij} (\mathbf{\beta}^\nu (\mathbf{\theta})).
\]

Combining (C.2.2), (C.2.3), (C.2.4), and (C.1.2), the gradient vector becomes:

\[
\mathbf{g}_N^S (\mathbf{\theta}) = \frac{1}{|N|} \sum_{i \in N} \frac{1}{|S_i|} \sum_{\nu \in S_i} \{ L_{ij} (\mathbf{\beta}^\nu (\mathbf{\theta})) \nabla \mathbf{\beta}^\nu (\mathbf{\theta}) \bar{\mathbf{x}}_i \mathbf{r}_{ij} (\mathbf{\beta}^\nu (\mathbf{\theta})) \}.
\]

In the above gradient calculation, the derivative of an instance of the relative importance vector \(\mathbf{\beta}^\nu (\mathbf{\theta})\) with respect to \(\mathbf{\theta}\) is required. This depends on the distribution of \(\mathbf{\beta}\).

Suppose the relative importance vector \(\mathbf{\beta} \in \mathbb{R}^K\) is distributed normal with the means vector \(\mu\) and the covariance matrix \(\Sigma\):

\[
\mathbf{\beta} \sim N(\mu, \Sigma).
\]

The covariance matrix can be rewritten as:

\[
\Sigma = Q'Q,
\]

where \(Q'Q\) is the upper triangular Cholesky factorization of \(\Sigma\). In this case, the parameter vector is given by:

\[
\mathbf{\theta} = \{ \mu_1, \mu_2, \ldots, \mu_K, q_{11}, q_{12}, \ldots, q_{1K}, q_{22}, q_{23}, \ldots, q_{2K}, \ldots, q_{KK} \},
\]

where \(q_{ij}\) is an element of the upper triangular matrix \(Q\).

Using the parameter vector, the instances of the relative importance vector can be generated by:

\[
\mathbf{\beta}^\nu_k (\mathbf{\theta}) = \mu_k + \sum_{j=k}^K \varepsilon_j^\nu q_{kj}, \forall k = 1, \ldots, K,
\]

104
where $\xi_k$ is an instance of standard normal distribution:

$$\nabla \beta^\nu (\theta) = I_{K \times K}$$

$$\xi_1' 0 \ldots 0$$

$$\xi_2' 0 \ldots 0$$

$$\vdots 0 \ldots 0$$

$$\xi_K' 0 \ldots 0$$

$$0 \xi_2 \ldots 0$$

$$0 \vdots \ldots 0$$

$$0 \xi_K \ldots 0$$

$$\vdots \ldots \vdots$$

$$0 0 \ldots \xi_K$$

The Hessian of the simulated log-likelihood function is calculated as follows:

$$H^S_N(\theta) := \nabla^2 SLL^S_N(\theta)$$

$$= \frac{1}{|N|} \sum_{i\in N} \left\{ \frac{1}{SP_{ij}^S(\theta)} \nabla^2 SP_{ij}^S(\theta) - \frac{1}{(SP_{ij}^S(\theta))^2} \nabla SP_{ij}^S(\theta) (\nabla SP_{ij}^S(\theta))^T \right\}.$$  \hspace{1cm} (C.2.6)

In the above calculations, the second derivative of the simulated choice probabilities and the conditional choice probabilities are required:

$$\nabla^2 SP_{ij}^S(\theta) = \frac{1}{|S_i|} \sum_{\nu \in S_i} \nabla^2 L_{ij}(\beta^\nu(\theta)).$$ \hspace{1cm} (C.2.7)

The second derivative of the conditional choice probabilities are calculated as follows:

$$\nabla^2 L_{ij}(\beta^\nu(\theta)) = \nabla \beta^\nu(\theta) \nabla^2 \theta (\nabla L_{ij}(\beta^\nu(\theta)))' + \nabla^2 \beta^\nu(\theta) \nabla L_{ij}(\beta^\nu(\theta))$$

$$= \nabla \beta^\nu(\theta) \nabla^2 L_{ij}(\beta^\nu(\theta)) (\nabla \beta^\nu(\theta))'$$

$$+ \nabla^2 \beta^\nu(\theta) \nabla L_{ij}(\beta^\nu(\theta)).$$ \hspace{1cm} (C.2.8)
Substituting the equations (C.2.7) & (C.2.8) into equation (C.2.6), the Hessian matrix can be rewritten as:

$$H^S_N(\theta) = \frac{1}{|N|} \sum_{i \in N} \left\{ \frac{1}{S_P^{S_i}(\theta)} \left[ \frac{1}{|S_i|} \sum_{\nu \in S_i} \left\{ \nabla \beta^\nu(\theta) \nabla^2 L_{ij}(\beta^\nu(\theta)) (\nabla \beta^\nu(\theta))^\prime \right\} 
\right.
+ \nabla^2 \beta^\nu(\theta) \nabla L_{ij}(\beta^\nu(\theta)) \right\}
\right.
- \frac{1}{(S_P^{S_i}(\theta))^2} \left[ \frac{1}{|S_i|} \sum_{\nu \in S_i} \nabla \beta^\nu(\theta) \nabla L_{ij}(\beta^\nu(\theta)) \right]
\right.
\left. \left( \frac{1}{|S_i|} \sum_{\nu \in S_i} \nabla \beta^\nu(\theta) \nabla L_{ij}(\beta^\nu(\theta)) \right)^\prime \right\}. \quad (C.2.9)
APPENDIX D

DATA SOURCES

D.1 Sales Transaction Data

The sales transaction dataset consists of various fields, some of which are listed below. The definitions of these fields [33] are briefly summarized.

- Nameplate/Make: The brand name or make of the vehicle (e.g., Chevrolet, Toyota).
- Model: The vehicle model (e.g., Cobalt, Civic, Corolla).
- Model Year: The vehicle model year (e.g., 2004).
- Segment: The vehicle classification (e.g., Compact, Large).
- Manufacturer: Manufacturer of the vehicle (e.g., Ford, GM, DaimlerChrysler).
- Series/Trim Level: Classification of the vehicle within the same model (e.g., LS, LT).
- Region: The geographic area in which the vehicle was sold.
- Transaction Type: Classification of the sale as cash, finance or lease.
- Sale Amount/Vehicle Price: The price that the customer pays for the vehicle and accessories and options.
- Total Down Payment: The total amount of cash and equity applied as a down payment.
• Rebate: The cash amount given to the customer by the manufacturer and/or retailer and used for the purchase or lease of a vehicle on the vehicle.

• Amount Financed: Portion of the purchase price funded by the lender.

• APR (Annual Percentage Rate): Annualized percentage rate paid by a customer for finance and lease transactions.

• Monthly Payment: The amount owed to the lender/lessor each month.

• Term Length: The number of months that a customer will make finance or lease payments on the vehicle.

• Capital Reduction: The prepayment of a leased vehicle’s depreciation, made by the customer at the time of the lease transaction.

• Net Capital Cost: Portion of the purchase price funded by the lender for lease transactions.

• Residual Amount: The dollar value of the vehicle at the end of the lease, estimated at the beginning of the lease.

• Trade In Amount: The dollar amount the retailer allows the customer for the trade-in vehicle.

• Trade In Payoff: The amount the customer still owes on his previously-financed trade in.

• Trade In Equity: The customer’s equity in the trade-in vehicle:

\[
\text{Trade In Equity} := \text{Trade In Amount} - \text{Trade In Payoff}.
\]

• Actual Cash Value

• Days-To-Turn: The number of days between the date that the vehicle arrived at the retailer and the date that the vehicle was purchased or leased.
D.2 Vehicle Attributes Data

- Fuel Economy City: is the fuel economy of the vehicle measured in miles per gallon in a city or urban driving cycle.

- Fuel Economy Highway: is the fuel economy of the vehicle measured in miles per gallon in a highway or extra-urban cycle.

- Fuel Economy Combined: is the average of the two fuel economy figures (Fuel Economy City, Fuel Economy Highway).

- Residual Percent After 2 years: is the percent ratio of the vehicle's value after 2 years of depreciation to the vehicle's initial value.

- Interior Passenger Volume: is an estimate of the size of the interior passenger compartment of the vehicle.

- Horsepower

- Length

- Width
REFERENCES


VITA

Deniz Dogan was born in Istanbul, Turkey on August 15, 1977. He received his B.S. degree in Industrial Engineering and Computer Engineering and M.S. degree in Industrial Engineering from Bogazici University, Istanbul, Turkey. During summers of 2004, 2005, and 2006, he was employed by General Motors Corporation where he received the Most Valuable Colleague Award. His doctoral research concentrated on discrete choice demand models and optimization. After receiving a PhD in Industrial Engineering from the Georgia Institute of Technology, he joined J.D. Power and Associates as a senior scientific programmer in May 2007.