ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Maximum Simulated Likelihood Estimation of the Mixed Logit Model

Submitted By:
Yannick van Gilst - *435273*
Supervisor: A. Castelein
Second assessor: C.D. van Oosterom

July 7, 2019

Abstract

This thesis investigates the estimation of the parameters of a mixed logit model. We use maximum simulated likelihood estimation where we compare two methods to evaluate the integrals in the simulated log likelihood of the mixed logit model. The first method is the pseudo-random Monte Carlo (PMC) method, which uses pseudo-random numbers to evaluate the integrals. The second method is the quasi-random Monte Carlo (QMC) method, which uses Halton draws to evaluate the integrals. We compare the performance of both methods with numerical experiments using data about ketchup brand choices. We find that the QMC method provides better accuracy, although the difference with the PMC method is small. We also considered a latent class mixed logit (LCML) model as an extension of the mixed logit model. However, based on the Bayesian information criterion we found that for our data set the mixed logit model is preferred over the LCML model.

ERASMUS UNIVERSITEIT ROTTERDAM

Contents

1 Introduction

The mixed logit model is a discrete choice model that, for example, can be used to analyze preferences of individuals or to predict behaviour. This model is a lot more flexible than the multinomial logit (MNL) model. The MNL model can be restrictive in modeling behaviour because e.g. individuals are assumed to have the same preferences. The mixed logit model, on the other hand, accounts for unobserved preference heterogeneity, which means that some individuals might value certain variables differently than others.

Another model that also accounts for unobserved preference heterogeneity is the latent clas logit model. In this model, it is assumed that there is a specific amount of classes and within such a class the individuals are assumed to be homogeneous. However, it could still be possible to have some heterogeneity within a class. To reveal additional dimensions of preference heterogeneity within a class, we construct a latent class mixed logit (LCML) model as described in Greene and Hensher (2013).

The estimation of the mixed logit and LCML model is however less straightforward than for the MNL model. It is therefore very useful to know what the best and most efficient estimation methods are for these models. We could estimate the parameters of these models by optimizing the corresponding log likelihood function. However, this log likelihood function consists of integrals and the number of integrals depends on the number of independent variables. One could imagine that, if the number of independent variables increases, optimizing the log likelihood function becomes really hard and time-consuming. Fortunately, we can use simulation methods to evaluate the integrals in the log likelihood. With these simulation methods, we are able to approximate the real log-likelihood using the simulated log-likelihood. The simulation methods that we investigate in this thesis, are the pseudo-random Monte Carlo (PMC) method and the Quasi-random Monte Carlo (QMC) method. For the first method, we use pseudo-random draws and for the latter method, we construct Halton draws.

In the first part of this thesis, we review the methods based on accuracy and estimation speed using numerical experiments. The data, we use to investigate the estimation methods, contains information about ketchup brand choices from 300 households with a total of 2798 observations in Springfield, Missouri (Jain et al., 1994). In the second part of this thesis, we use the method that seems to provide the best accuracy in three different empirical applications with our data set. We consider here a mixed logit model with either normal distributed dependent or independent coefficients and the latent class logit model with normal distributed independent coefficients.

We find, based on the numerical experiments, that the accuracy provided by the PMC and the QMC method improves if the number of draws increases. We find that the PMC method with 25 draws provides better accuracy than the QMC method with 25 draws. If the number of draws is equal to 50 or more, the QMC method provides, depending on the number of draws, about 1 to 5% better accuracy compared to the PMC method. These results are not consistent with earlier research about this topic, like Bhat (2001), where they concluded that the QMC method provides a considerably better accuracy with much fewer draws than the PMC method needs to achieve the same accuracy.

We find, for our data set, that the mixed logit model, where the coefficients are assumed to be independent and follow the normal distribution, is preferred over the mixed logit model where the coefficients are assumed to be correlated and follow the multivariate normal distribution. The mixed logit model with independent, normal distributed coefficients is also preferred over the latent class mixed logit model with independent, normal distributed coefficients.

Many researchers can potentially benefit from the results of this thesis. Unobserved heterogeneity is a frequent occurring phenomenon and it is therefore essential to have a model that is able to best capture this. With the results of this thesis, it becomes clear what the best approach is to estimate these models. Hence, researchers are less prone to making errors and they could save time by using the most efficient and fastest estimation method.

The structure of the thesis is as follows. Section 2 gives an overview of some of the existing literature on this topic. Section 3 explains the data we use in our research. Section 4 describes the methods in our research in more detail and Section 5 gives the results. Finally, in Section 6 we give a conclusion.

2 Literature

The mixed logit model is an often used model and is being used for quite some time now. Therefore, much literature is available on this topic. The mixed logit model was introduced in Boyd and Mellman (1980) and Cardell and Dunbar (1980). Due to an increase in computer power since this time period, this model became more popular and researchers were able to make better use of this model because simulation went a lot faster.

The starting point of our research is Bhat (2001). In this paper, the author compares three different methods to evaluate the integrals in the likelihood function of a mixed logit model. The methods he compares are the polynomial-based cubature (PBC) method, the pseudo-random Monte Carlo (PMC) method and the quasi-random Monte Carlo (QMC) method. The author uses data on intercity travel mode choices. The independent variables he is using in the model comes from actual field data. The choice process, however, is generated by simulating normal distributed random coefficients. The author finds that the QMC method provides better accuracy in a lot less computation time than the PBC and PMC method. Furthermore, the QMC method needs much fewer draws than the PMC method to reach the same accuracy. Just like Bhat (2001), we also compare the PMC and the QMC method. The difference, however, is that we also consider coefficients that are assumed to be multivariate normal distributed instead of only independent normal distributed coefficients.

Train (2000) find similar results as in Bhat (2001) in another application. However, the author also warns us that there is much that remains to be investigated when using Halton draws for estimating a mixed logit model. The author explains that there might be a relationship between the number of draws used per observation and the primes that are used to create the Halton draws. One of the questions that stays unsolved here is if this is desirable or not. Bhat (2003) describes a solution for such a relationship, namely scrambled Halton draws. The author finds that the use of scrambled Halton draws improves the efficiency and the estimation speed of the QMC method compared to the use of regular Halton draws in a model with high dimensions (i.e. ≥ 10). However, we only consider models with low dimensions (i.e. ≤ 10) and therefore we do not need to use scrambled Halton draws.

Hensher and Greene (2003) discusses the mixed logit model and the experiences of active researchers working on the mixed logit model at the time of publishing. They found that the main challenge the researchers face with the mixed logit model comes from the data they use. The mixed logit model demands better quality data than the multinomial logit model. They argue that better quality data is needed to better understand the true behavioural variability between individuals.

We also consider the latent class mixed logit (LCML) model, described in Greene and Hensher (2013). This is an extension to the latent class model. The LCML model also allows preference heterogeneity within each class in contrast to a latent class model where preference homogeneity is usually assumed within a class. They argue that the gains of allowing random parameters in a latent class model, probably depends on the data-set. Therefore, they conclude that more research should be done on this model with more data sets. Using our data set of ketchup brand choices, we can thus contribute to their research.

3 Data

The data we use for our research is panel data describing purchases of ketchup of households in Springfield, Missouri in a time period of about two years (Jain et al., 1994). We have information about 300 households with a total of 2798 observations. The data contains information about two brands of ketchup, namely Heinz and Hunts. Heinz sells three different sizes of 41, 32 and 28 ounces and Hunts is selling only one size of 32 ounces. We have information on all four products about the price, if the product is on display and if there is a newspaper feature advertisement for the product at the time of purchase.

Table 1 summarizes the characteristics of the dependent and explanatory variables for the four products. In the table, we see that Heinz (32 oz.) has the highest market share of 52.1%. This product has also the lowest average price and was on display the most of all four products. Heinz (28 oz.) has the most feature advertisements in the newspaper and is on display and has a feature at the same time the most of all four products.

Furthermore, the data shows us that the average price of Heinz (41 oz.) has a

Table 1: *Characteristics of the dependent variable and explanatory variables for two ketchup brands with varying sizes*

Variables	Heinz (41 oz.)	Heinz (32 oz.)	Heinz (28 oz.)	Hunts (32 oz.)
Choice percentage	0.065	0.521	0.304	0.110
Average price (US$/oz.)	0.113	0.098	0.154	0.105
% display only[a]	2.14	8.61	6.15	3.54
% feature only[b]	3.14	5.22	5.36	3.65
% feature and display[c]	0.11	1.32	1.50	0.93

Notes:
[a] Percentage of purchase occasions when a brand was on display only
[b] Percentage of purchase occasions when a brand was featured only.
[c] Percentage of purchase occasions when a brand was on display and featured.

standard deviation of about 0.011. The minimum average price of Heinz (41 oz.) is \$0.049 and the maximum \$0.176. The standard deviation of the average price of Heinz (32 oz.) is about 0.017 and has a minimum and maximum average price of respectively \$0.009 and \$0.134. The standard deviation of the average price of Heinz (28 oz.) is about 0.027 and has a minimum and maximum average price of respectively \$0.004 and \$0.429. Lastly, the standard deviation of the average price of Hunts (32 oz.) is about 0.017 and has a minimum and maximum average price of respectively \$0.009 and \$0.272.

4 Methodology

4.1 Mixed logit model

The mixed logit model is a discrete choice model that, as mentioned before, accounts for unobserved preference heterogeneity. The notations and derivations used in this section come from Train (2009). The general form of the mixed logit model can be derived from the utility. The utility function for individual i of alternative j at choice time t is given as:

$$U_{ijt} = \beta_i' x_{ijt} + \epsilon_{ijt} \tag{4.1}$$

where x_{ijt} is a $(K \times 1)$ vector of K observed independent variables for individual i for alternative j at choice time t. β_i is a vector of K coefficients specific to individual i with density $f(\beta_i|\theta)$, where θ represents the parameters of this density function . Furthermore, ϵ_{ijt} is a random term which is i.i.d. type 1 extreme value distributed. Moreover, we have that $i \in \{1, ..., N\}$, $j \in \{1, ..., J\}$ and $t \in \{1, ..., T_i\}$. Individual i then chooses alternative j at choice time t if and only if $U_{ijt} > U_{ilt}$ $\forall l \neq j$.

The mixed logit formula, which gives us the unconditional probability that individual i chooses alternative j at choice time t, is then given as:

$$P_{ijt}(\theta) = \int L_{ijt}(\beta_i) f(\beta_i|\theta) d\beta_i \tag{4.2}$$

where $L_{ijt}(\beta_i)$ is the logit probability conditional on β_i and is given as:

$$L_{ijt}(\beta_i) = \frac{exp(\beta_i' x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i' x_{ilt})} \tag{4.3}$$

such that the mixed logit probability can be written as:

$$P_{ijt}(\theta) = \int \frac{exp(\beta_i' x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i' x_{ilt})} f(\beta_i|\theta) d\beta_i \tag{4.4}$$

The mixing distribution $f(\beta_i|\theta)$ can be either discrete or continuous. The normal, log-normal, triangular or uniform distribution are often used distributions for the mixed logit model. In our research, we only focus on the univariate and multivariate normal distribution.

4.1.1 Parameter estimation

To get the estimates of the parameters of the mixed logit model, one can use maximimum likelihood estimation. The likelihood contribution for individual i may be written as:

$$\ell_i = \int \prod_{t=1}^{T_i} \prod_{j=1}^{J} {L_{ijt}}^{\delta_{ijt}} f(\beta_i|\theta) d\beta_i = \int \prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[\frac{exp(\beta_i' x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i' x_{ilt})} \right]^{\delta_{ijt}} f(\beta_i|\theta) d\beta_i \tag{4.5}$$

where δ_{ijt} is a dummy equal to 1 if individual i chooses alternative j at choice time t and zero otherwise. The likelihood function is then defined as:

$$L = \prod_{i=1}^{N} \ell_i = \prod_{i=1}^{N} \int \prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[\frac{exp(\beta_i^{'} x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i^{'} x_{ilt})} \right]^{\delta_{ijt}} f(\beta_i|\theta) d\beta_i \tag{4.6}$$

and the log likelihood can then be written as:

$$\mathscr{L} = \sum_{i=1}^{N} ln \left(\int \prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[\frac{exp(\beta_i^{'} x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i^{'} x_{ilt})} \right]^{\delta_{ijt}} f(\beta_i|\theta) d\beta_i \right) \tag{4.7}$$

Due to the fact that the log likelihood in (4.7) contains multiple integrals, it can be very hard and time consuming to optimize this expression. Therefore, we use maximum simulated likelihood estimation (MSLE) to get the parameter estimates. The simulated log likelihood (SLL) is given by:

$$SLL = \sum_{i=1}^{N} ln \left[\frac{1}{R} \sum_{r=1}^{R} \left(\prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[\frac{exp(\beta_i^{r'} x_{ijt})}{\sum_{l=1}^{L} exp(\beta_i^{r'} x_{ilt})} \right]^{\delta_{ijt}} \right) \right] \tag{4.8}$$

with R equal to the number of draws used in the optimization of this function.

The pseudo-random Monte Carlo (PMC) method and the quasi-random Monte Carlo (QMC) method can be used to generate the R draws $\{\beta_i^r\}_{r=1}^R$ for every $i \in \{1, ..., N\}$. If we let β_i be independent and follow the normal distribution, such that $\beta_{i,k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$, we can write:

$$\beta_{i,k}^r = \mu_k + \sigma_k s_{i,k}^r \tag{4.9}$$

where $s_{i,k}^r$ is a standard normal variate, generated with the PMC or QMC method.

If we allow the coefficients to be correlated, such that the coefficients follow the multivariate normal distribution, we have that $\beta_i \sim \mathcal{N}(\mu, \Sigma)$. This means that:

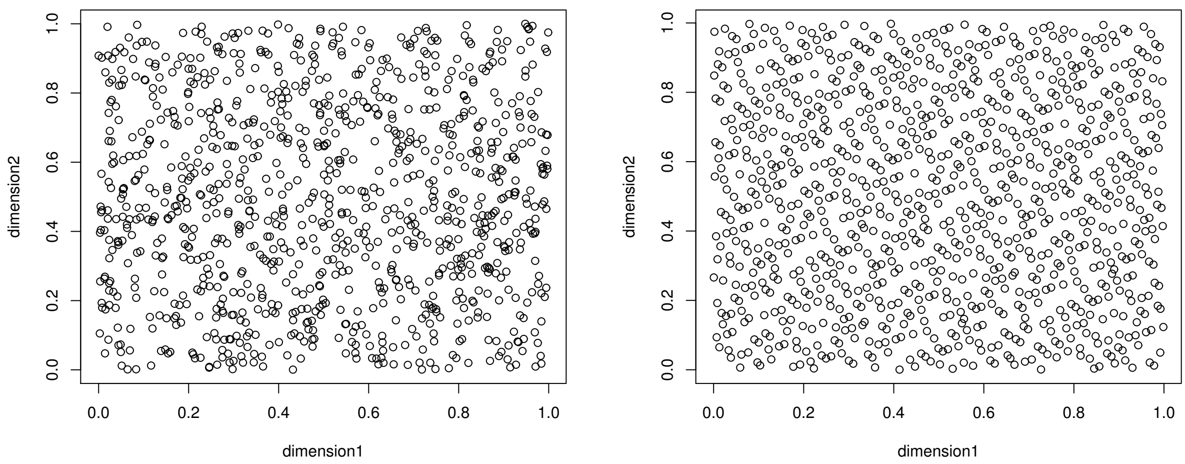
$$\beta_i^r = \mu + L s_i^r \tag{4.10}$$

where L is the Choleski factor of Σ, such that $LL^{'} = \Sigma$ and s_i^r is a $(K \times 1)$ vector of standard normal variates.

In both situations, we can optimize the SLL using the BFGS method, such that we get the estimates $\hat{\mu}_k$ and $\hat{\sigma}_k$ if the coefficients are independent and $\hat{\mu}$ and $\hat{\Sigma}$ if the coefficients are correlated.

The difference between the PMC and the QMC method lies in the way how matrix S_i is generated. The PMC method directly generates a $(R \times K)$ matrix S_i of standard normal random numbers for every individual i. We then optimize (4.8) where $s^r_{i,k}$ is the element at position (r, k) of matrix S_i specific to individual i.

The QMC method generates matrix S_i for individual i using Halton draws. By taking the inverse cumulative normal of a Halton draw, one can obtain a standard normal variate needed in matrix S_i. Halton draws are said to span the M-dimensional unit cube in a more equal way than random draws from the uniform distribution with parameters 0 and 1. One can see in Figure 1 that the pseudo-random sequence, generated from a



(a) Pseudo-random sequence

(b) Halton sequence

Figure 1: 1000 draws from a pseudo-random sequence and from a Halton sequence in two dimensions

uniform distribution with parameters 0 and 1, are more clumped together and provide a less even coverage along the two-dimensional space than the Halton sequence.

We can create a Halton sequence by first choosing a prime p. Then, we can iteratively create a Halton sequence, where the sequence is defined at iteration $t+1$ as:

$$h_{t+1} = \{h_t, h_t + \frac{1}{p^t}, h_t + \frac{2}{p^t}, ..., h_t + \frac{p-1}{p^t}\} \tag{4.11}$$

with $h_1 = \{0\}$. We are now able to create a matrix S_i for every individual i using Halton draws. We need K different primes to create K Halton sequences with each $N * R + c$ Halton draws, where c is the maximum value of the K primes. We then discard the first c elements of all Halton sequences to prevent that there is correlation between the K sequences. We then merge all Halton sequences to create one matrix and we call this the Halton matrix which has dimension $(N * R) \times K$. Now, we can make matrix S_i for every individual i if we split the Halton matrix into N different submatrices with dimension $R \times K$ and taking the inverse cumulative normal of every element in these matrices. Now that we have a matrix S_i for every individual, we can again optimize the SLL.

Maximimum simulated likelihood (MSL) is consistent, asymptotically normal and efficient and asymptotically equivalent to maximum likelihood if the number of draws R rises faster than sample size S. If R rises slower than sample size S, MSL is still consistent but not asymptotically normal. Finally, if R is fixed, the simulation bias will rise as S rises and therefore MSL is inconsistent in this situation. This means that if the number of draws rises faster than the sample size, the MSL estimator $\hat{\theta}$ is distributed as follows:

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, -\mathbf{H}^{-1}/S) \tag{4.12}$$

where θ are the true parameters and the information matrix:

$$-\mathbf{H} = -E\left(\frac{\partial^2 SLL(\theta)}{\partial\theta\partial\theta'}\right) \tag{4.13}$$

with $SLL(\theta)$ equal to the SLL evaluated at parameters θ. The standard error of the jth estimated parameter $\hat{\theta}_j$ can then be calculated as $\sqrt{-\hat{\mathbf{h}}_{jj}^{-1}/S}$ where $\hat{\mathbf{h}}_{jj}$ is the jth diagonal element of the information matrix evaluated at estimated parameters $\hat{\theta}$.

4.1.2 Parameter interpretation

Now that we estimated the parameters, the question remains how to interpret these results. If we assume that $\beta_{i,k}$ is independent and follows the normal distribution, we get from the estimation of the mixed logit model that $\beta_{i,k} \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$ for all $k \in \{1, ..., K\}$. With the cumulative distribution function (CDF) of the normal distribution with parameters $\hat{\mu}_k$ and $\hat{\sigma}_k$, we can calculate the probability that $\hat{\beta}_{i,k}$ has a value less than zero. If this probability is equal to p, we know that the share of individuals, for which an increase of variable k for alternative j has a negative effect on the probability of choosing alternative j, is equal to p. Hence, the share of individuals, for which an increase of variable k for alternative j has a positive effect on the probability of choosing alternative j, is equal to $p - 1$. If we assume that the coefficients are correlated and follow the multivariate normal distribution, we can find similar results if we use the CDF of the multivariate normal distribution. We can then calculate $\mathbb{P}(\beta_i \leq b)$, with b a vector of values. Since we are interested in the probability that a parameter is less than zero, we set the corresponding element in vector b to zero and the remaining elements to $+\infty$. Likewise, we can also calculate the probability that multiple parameters are less than zero. These probabilities are numerically estimated as described in Botev (2017).

4.2 Latent class mixed logit model

As an extension to the latent class model, Greene and Hensher (2013) came up with the latent class mixed logit model. This model reveals additional dimensions of preference heterogeneity. In a latent class model, as described in Greene and Hensher (2003), we assume that there are Q groups in the population. Within such a group, the individuals are homogeneous and share the same interests, but the groups are different from each other. In a latent class mixed logit model, however, it is also possible to have heterogeneity within a group.

It is unknown to the researcher in which class an individual belongs. However, we do know that the probability that an individual is in class q is given by:

$$P(class = q) = \pi_q, \qquad q = 1, ..., Q \tag{4.14}$$

with $0 \leq \pi_q \leq 1$ $\forall q$ and $\sum_q \pi_q = 1$.

The probability that individual i chooses alternative j at choice time t is now given as:

$$P_{ijt}(\theta) = \sum_{q=1}^{Q} \pi_q \int \frac{exp(\beta_{iq}' x_{ijt})}{\sum_{l=1}^{L} exp(\beta_{iq}' x_{ilt})} f(\beta_{iq}|\theta) d\beta_{iq} \tag{4.15}$$

where $f(\beta_{iq}|\theta)$ is the density function and θ represents the parameters of this density.

4.2.1 Parameter estimation

To estimate the parameters of this model, we again use MSLE and the SLL is now given as:

$$SLL = \sum_{i=1}^{N} ln \left[\sum_{q=1}^{Q} \pi_q \frac{1}{R} \sum_{r=1}^{R} \left(\prod_{t=1}^{T_i} \prod_{j=1}^{J} \left[\frac{exp(\beta_{iq}^{r\prime} x_{ijt})}{\sum_{l=1}^{L} exp(\beta_{iq}^{r\prime} x_{ilt})} \right]^{\delta_{ijt}} \right) \right] \tag{4.16}$$

with δ_{ijt} again a dummy equal to 1 if individual i chooses alternative j at choice time t and zero otherwise and R is equal to the number of draws used in the optimization. Furthermore, we now only assume that β_{iq} follows the normal distribution for every $i \in \{1, ..., N\}$ and $q \in \{1, ..., Q\}$, such that $\beta_{iq,k} \sim \mathcal{N}(\mu_{q,k}, \sigma_{q,k}^2)$. Thus, we can write:

$$\beta_{iq}^r = \mu_{q,k} + \sigma_{q,k} s_{i,k}^r \tag{4.17}$$

where $s_{i,k}^r$ is a standard normal variate. With the PMC or QMC method, we can generate the R draws $\{\beta_{iq}^r\}_{r=1}^R$ for every $i \in \{1, ..., N\}$ and $q \in \{1, ..., Q\}$, in a similar way as for the mixed logit model. Optimizing (4.16) then gives us the parameter estimates $\hat{\pi}_q$, $\hat{\mu}_{q,k}$ and $\hat{\sigma}_{q,k}$ for every $j \in \{1, ..., J\}$, $q \in \{1, ..., Q\}$ and $k \in \{1, ..., K\}$. The standard errors can again be calculated as described in subsection 4.1.1

Since it is unknown what the optimal number of classes should be in the model, we have to estimate the parameters with a different number of classes. Based on the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), we then choose what the optimal number of classes should be.

4.2.2 Parameter interpretation

The interpretation of the parameters for the LCML model is similar to the interpretation of the parameters of the mixed logit model. From the estimation of the model, we now find that $\beta_{iq,k} \sim \mathcal{N}(\hat{\mu}_{q,k}, \hat{\sigma}^2_{q,k})$ for all $k \in \{1, ..., K\}$ and $q \in \{1, ..., Q\}$. Again, with the cumulative distribution function (CDF) of the normal distribution with parameters $\hat{\mu}_{q,k}$ and $\hat{\sigma}_{q,k}$, we can calculate the probability that $\beta_{iq,k}$ has a value less than zero. If this probability is equal to p, we know that the share of individuals in class q, for which an increase of variable k for alternative j has negative effect on the probability of choosing alternative j, is equal to p. Hence, the share of individuals in class q, for which an increase of variable k for alternative j has a positive effect on the probability of choosing alternative j, is equal to $p - 1$.

5 | Results

5.1 Comparison of the PMC and QMC method

To compare the performance of the PMC and the QMC method on the estimation of the mixed logit model, we do numerical experiments. To do this, we use the data as described in Section 3. However, we now use simulated choices instead of the actual choices made by the individuals. To simulate these choices, we calculate the utilities for every choice alternative per observation for every individual based on (4.1). The variables that we use to calculate the utility are price, display (i.e. dummy equal to 1 if the product is on display) and feature (i.e. dummy equal to 1 if the product has newspaper feature).

The coefficients for each individual are here assumed to be independent and are drawn from a normal distribution. The coefficient for price is drawn for every individual from a normal distribution with mean -1.5 and standard deviation 1. The coefficient for display is drawn with mean 1 and standard deviation 0.75 and the coefficient for feature is also drawn with mean 1 and standard deviation 0.75. Furthermore, the alternative specific constants are also drawn from a normal distribution, except the alternative specific constant for Hunts (32 oz.) which is set to zero for identification. The mean and standard deviation of the intercept for Heinz (41 oz.) are 1.25 and 0.9 respectively. The mean and standard deviation of the intercept for Heinz (32 oz.) are 1.5 and 1 respectively. The mean and standard deviation of the intercept for Heinz (28 oz.) are 2.5 and 1.8 respectively. The means are based on the parameter estimates when estimating the parameters of a conditional logit model with the data with the true choices. The standard deviations are approximately three-fourths of the corresponding means. Finally, the error terms are generated from the standard type I extreme value distribution.

The utilities for every individual for every alternative at every choice time are then calculated based on these randomly drawn coefficients and the choice alternative with the highest utility is then set as the chosen alternative.

We can now compare the performance of the PMC and QMC method based on the simulated choices. This is done by estimating the parameters of the mixed logit model using 25, 50, 100, 250 and 500 random draws per individual for the PMC method and with the same number of Halton draws per individual for the QMC method. For every number of draws, we estimate the parameters a hundred times using a different matrix S_i for every individual each estimation. For the PMC method, this is done by generating a new matrix S_i by generating its elements from a standard normal distribution. To obtain a different matrix S_i using the QMC method, we use a different combination without replacement of the first six primes to create Halton draws and then generating matrix S_i each estimation. The SLL was programmed in C++ and using the RCPP package then optimized in **R**. The optimization was done on a computer with an Intel Core i7 3.20 GHz processor with 16GB of RAM.

The performance is based on the proximity of the estimated parameters to the true parameters. This is done by evaluating the root mean square error (RMSE) of the parameters and is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2}{M}} \tag{5.1}$$

where $\hat{\theta}_m$ is the estimated parameter of the mth estimation, θ is the true parameter and M is the number of times the parameter is estimated, which in our case equal to hundred. The RMSE for every parameter for every different number of draws for both methods can be found in A.1. In Table 2, the sum of all RMSE's for both methods is given, which makes it easier to interpret the results. Furthermore, the average time to convergence is also given.

Table 2: *RMSE for the pseudo-random Monte Carlo method and Quasi-random Monte Carlo method.*

	Pseudo-random Monte Carlo method					Quasi-random Monte Carlo method				
	Number of draws					Number of draws				
	25	50	100	250	500	25	50	100	250	500
RMSE	3.882	3.610	3.289	2.997	2.880	3.900	3.575	3.136	2.900	2.773
Average time to convergence (min.)	0.62	1.20	2.18	5.26	9.88	0.62	1.19	2.20	4.97	9.64

Table 2 shows us that an increase in the number of draws leads to a lower RMSE for both estimation methods. Furthermore, the QMC method always has a lower RMSE compared to the PMC method if the number of draws is the same, except if the number of draws is equal to 25. However, the differences are small. We find that if the number of draws is 50 or larger, the QMC method results in a RMSE which is about 1 to 5% lower than the PMC method. Furthermore, the time to convergence is similar for both methods if the same number of draws is used.

5.2 Empirical applications

5.2.1 Mixed logit model with independent normal coefficients

To get a better insight into the mixed logit model, we consider the original data about ketchup brand choices. It is now assumed that the coefficients are independent and follow the normal distribution. We estimate the parameters of this model using the QMC method with 500 Halton draws per individual since this method resulted in the lowest RMSE according to our numerical experiment. Of course, adding more draws per individual will probably decrease the RMSE even more. However, according to theory, using 500 Halton draws will result in asymptotically consistent and efficient estimators for our data and therefore using 500 Halton draws is sufficient. The parameter estimates can be found in Table 3.

Table 3: *Parameter estimates.*

Variables	μ	σ
Intercepts		
Heinz(41 oz.)	1.948***	0.641
	(0.183)	(0.394)
Heinz(32 oz.)	1.734***	1.747***
	(0.144)	(0.075)
Heinz(28 oz.)	3.204***	1.263***
	(0.165)	(0.110)
Marketing variables		
Display	1.119***	0.750*
	(0.145)	(0.420)
Feature	1.277***	0.751*
	(0.167)	(0.440)
Price	-2.097***	1.152***
	(0.117)	(0.089)
SLL	-2082.4	
BIC	4260.0	
Time to convergence (min.)	11.9	

Notes:
*** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.
The standard errors are given in parentheses.
Hunts(32 oz.) is used as base brand.

First, we examine the marketing variables and how to interpret these results. Based on the estimates, we find that for roughly 97% of the individuals, an increase in the price of a product has a negative effect on the choice probability of this product. Hence, an increase of the price of a product has a positive effect on the choice probability of this product for 3% of individuals. Furthermore, for roughly 96% of individuals a newspaper feature for a product has a positive impact on the choice probability on choosing this product. Finally, a display for a product has positive impact on the choice probability of this product on about 93% of individuals

Secondly, we examine the intercepts, since they can give us information about the preference for a certain product if all marketing variables are equal for every product.

Since Hunts (32 oz.) is used as base brand, the preferences are relative to this brand. The probability that the intercept of Heinz (28 oz.) is lower than zero is smaller than 1%, meaning that less than 1% of individuals prefer Hunts (32 oz.) over Heinz (28 oz.) if their marketing variables are the same. With similar analysis, we find that roughly 16% of individuals prefers Hunts (32 oz.) over Heinz (32 oz.). Since the standard deviation of the intercept of Heinz (41 oz.) is not significantly different from zero at the 10% level, it seems that there always is a preference for Heinz (41 oz.) over Hunts (32 oz.).

5.2.2 Mixed logit model with multivariate normal coefficients

We now consider the mixed logit model, where the coefficients are assumed to be correlated and follow the multivariate normal distribution. We again estimate the parameters of this model with the QMC method with 500 Halton draws per individual for the same reasons as given before, such that we find $\beta_i \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$. The estimates $\hat{\mu}$ are given in Table 4 and given that $\hat{\Sigma} = \hat{L}\hat{L}'$, we find that:

$$\hat{\Sigma} = \begin{bmatrix} 4.174 & 4.057 & 4.483 & 0.732 & -0.087 & 0.369 \\ 4.057 & 6.627 & 3.665 & 1.917 & -0.595 & 0.281 \\ 4.483 & 3.665 & 5.693 & -0.051 & -0.190 & -0.355 \\ 0.732 & 1.917 & -0.051 & 1.190 & 0.115 & 0.459 \\ -0.087 & -0.595 & -0.190 & 0.115 & 0.619 & 0.137 \\ 0.369 & 0.281 & -0.355 & 0.459 & 0.137 & 1.365 \end{bmatrix}$$

with estimated $\hat{L}$ and corresponding standard errors given in A.2.

With these results, it is now possible to calculate the share of individuals whose brand choice is either positively or negatively affected by a certain marketing variable. We now find that a product on display positively affects the choice probability of this product for about 86% of individuals. Furthermore, a product with a newspaper feature positively affects the choice probability of this product for about 96% of individuals and an increase of the price of a product has a negative impact on the choice probability of this product for about 97% of individuals. These results are similar to what we find for the mixed logit model with independent coefficients.

By examining the intercepts, we can obtain information about the preference of a product if the marketing variables are the same for every product. Hunts (32 oz.) is

Table 4: *Parameter estimates.*

Variables	μ
Intercepts	
Heinz(41 oz.)	2.844***
	(0.271)
Heinz(32 oz.)	2.513***
	(0.243)
Heinz(28 oz.)	4.134***
	(0.269)
Marketing variables	
Display	1.162***
	(0.159)
Feature	1.347***
	(0.180)
Price	-2.177***
	(0.129)
SLL	-2028.3
BIC	4271.0
Time to convergence (min.)	44.8

Notes:
*** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.
The standard errors are given in parentheses.
Hunts(32 oz.) is used as base brand.

used as base brand and the preferences are therefore relative to this brand. We find that Heinz (41 oz.) is preferred over Hunts (32 oz.) by about 92% of individuals, Heinz (32 oz.) by about 84% and Heinz (28 oz.) by about 96% if all marketing variables are equal. These results are also similar to what we find for the mixed logit model with independent coefficients.

This model, however, results in a higher BIC than the model with independent normal distributed coefficients. This means that the model with independent coefficients is preferred over this model with correlated coefficients.

5.2.3 Latent class mixed logit model

We now consider the latent class mixed logit model on our data, where the coefficients are assumed to be independent and normal distributed. To see what the optimal number of classes in the model should be, we examine the BIC for the LCML with a different number of classes. The estimation is again done using the QMC method with 500 Halton draws per individual. We start by estimating the parameters of a model with two classes. Then, we estimate the parameters with three classes. If the BIC is lower for the model with three classes, we estimate the parameters with four classes and so forth. Table 5 shows these results.

Table 5: *Results for LCML models with different number of classes.*

	Number of classes		
	2	3	4
SLL	-2045.0	-2028.3	-2027.6
BIC	4288.4	4358.2	4460.0
Parameters[a]	25	38	51
Time to convergence (min.)	80.3	247.8	584.6

Notes:
[a] Number of parameters in the model to be estimated

The latent class mixed logit model with two classes has the lowest BIC, with a value of 4288.4. Table 6 shows the parameter estimates of this model. Based on these estimates, the effects of a change of a variable on the choice probability of a product can again be calculated in both classes. It seems, for example, that a product on display has no impact on the choice probability of this product for individuals in class 2, while it does have a positive effect on the choice probability for individuals in class 1. Furthermore, it seems that almost all individuals in class 1 prefer any size of Heinz ketchup over Hunts (32 oz.) if all marketing variables are the same. Individuals in class 2, on the other hand, seem to prefer Hunts (32 oz.) over Heinz (32 oz.) and are indifferent between Heinz (41 oz.) and Hunts (32 oz.) if the marketing variables are the same.

We can compare the BIC of the LCML model with the BIC of the mixed logit

Table 6: *Parameter estimates of LCML model with 2 classes.*

	Class			
	1		2	
Variables	μ	σ	μ	σ
Intercepts				
Heinz(41 oz.)	2.945***	0.447	0.023	0.726
	(0.260)	(0.852)	(0.558)	(1.113)
Heinz(32 oz.)	2.760***	1.704***	-0.511**	0.547
	(0.219)	(0.096)	(0.224)	(0.631)
Heinz(28 oz.)	4.053***	0.936***	1.560***	1.838***
	(0.239)	(0.177)	(0.472)	(0.216)
Marketing variables				
Display	1.428***	0.009	-0.117	0.009
	(0.164)	(19.277)	(0.362)	(24.522)
Feature	1.407***	0.564	1.017**	0.377
	(0.202)	(0.736)	(0.448)	(2.778)
Price	-1.932***	1.060***	-2.459***	0.011
	(0.143)	(0.116)	(0.265)	(16.877)
π	0.801		0.199	

Notes:
*** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level.
The standard errors are given in parentheses. Hunts(32 oz.) is used as base brand. π denotes the share of individuals in the corresponding class.

model with independent coefficients, which can be seen as a LCML model with just one class. The BIC of the mixed logit model is 4260.0, which is lower than the BIC of the LCML model with two classes. Hence, the mixed logit model fits the data better than the LCML model. Furthermore, the time to convergence is much more compared to the mixed logit model and every class that we add to the model drastically increases the time to convergence. This can be explained with the fact that we add thirteen new parameters to the model to be estimated if we increase the number of classes by one.

6 Conclusion

In this research, we evaluated the integrals of the log likelihood function of the mixed logit model with two different methods, i.e. the pseudo-random Monte Carlo and Quasi-random

Monte Carlo method. We compared both methods based on their accuracy using numerical experiments. Based on these numerical experiments, we find that the accuracy of both methods increases if the number of draws increases. Furthermore, we find that the QMC method has a better accuracy than the PMC method if the number of draws is the same, but larger than 50. The QMC method provides in this situation between 1 to 5% better accuracy.

In conclusion, we can say that although the QMC method provides a better accuracy, our results are not completely consistent with earlier research about this topic, like Bhat (2001). According to this paper, we should find that the QMC method provides considerably better accuracy with much fewer draws needed than the PMC method to obtain the same accuracy, which is not the case in our research. This means that the actual benefits of using the QMC method might depend on the data set and this should thus be investigated in more detail.

In the second part of this research, we considered three models with the data set about ketchup brand choices. We considered a mixed logit model with normal distributed independent coefficients, a mixed logit model with multivariate normal distributed coefficients and a latent class class mixed logit model with normal distributed independent coefficients. In all three applications, it seems that a product on display has a positive effect on the choice probability of this product for a majority of individuals. A newspaper feature also has a positive effect and an increase in price a negative effect on the choice probability of a product for the majority of individuals in all three models. Finally, we find that the mixed logit model with independent normal distributed coefficients is preferred based on the Bayesian Information Criterion.

References

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7):677–693.

Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. *Transportation Research Part B: Methodological*, 37(9):837–855.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148.

Boyd, J. H. and Mellman, R. E. (1980). The effect of fuel economy standards on the us automotive market: an hedonic demand analysis. *Transportation Research Part A: General*, 14(5-6):367–378.

Cardell, N. S. and Dunbar, F. C. (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14(5-6):423–434.

Greene, W. H. and Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8):681–698.

Greene, W. H. and Hensher, D. A. (2013). Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45(14):1897–1902.

Hensher, D. A. and Greene, W. H. (2003). The mixed logit model: the state of practice. *Transportation*, 30(2):133–176.

Jain, D. C., Vilcassim, N. J., and Chintagunta, P. K. (1994). A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics*, 12(3):317–328.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Train, K. (2000). Halton sequences for mixed logit.

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge university press.

A Appendix

A.1 RMSE for the PMC and QMC method

Table 7: *RMSE for parameters based on the PMC method.*

	Number of draws									
	25		50		100		250		500	
Variables	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Intercepts										
Heinz(41 oz.)	0.364	0.527	0.393	0.423	0.419	0.262	0.448	0.135	0.457	0.128
Heinz(32 oz.)	0.084	0.203	0.072	0.125	0.060	0.090	0.045	0.049	0.038	0.044
Heinz(28 oz.)	0.422	0.135	0.389	0.100	0.352	0.089	0.317	0.054	0.308	0.039
Marketing variables										
Display	0.214	0.660	0.236	0.687	0.253	0.658	0.261	0.664	0.264	0.675
Feature	0.033	0.631	0.054	0.604	0.067	0.560	0.076	0.490	0.080	0.402
Price	0.391	0.218	0.346	0.182	0.311	0.170	0.286	0.171	0.276	0.167

Table 8: *RMSE for parameters based on the QMC method.*

	Number of draws									
	25		50		100		250		500	
Variables	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Intercepts										
Heinz(41 oz.)	0.364	0.575	0.403	0.473	0.419	0.236	0.443	0.115	0.460	0.132
Heinz(32 oz.)	0.075	0.252	0.061	0.114	0.053	0.061	0.037	0.044	0.038	0.036
Heinz(28 oz.)	0.425	0.122	0.374	0.102	0.320	0.067	0.302	0.047	0.300	0.035
Marketing variables										
Display	0.229	0.659	0.246	0.659	0.258	0.650	0.265	0.657	0.267	0.660
Feature	0.044	0.581	0.064	0.556	0.072	0.547	0.081	0.465	0.084	0.324
Price	0.370	0.204	0.333	0.191	0.295	0.158	0.280	0.164	0.271	0.167

A.2 Estimation and standard errors of L

$$\hat{L} = \begin{bmatrix} -2.043*** & 0 & 0 & 0 & 0 & 0 \\ -1.986*** & -1.638*** & 0 & 0 & 0 & 0 \\ -2.194*** & 0.422*** & -0.836*** & 0 & 0 & 0 \\ -0.358 & -0.736*** & 0.630*** & 0.351 & 0 & 0 \\ 0.042 & 0.312 & 0.273 & 0.534* & -0.401 & 0 \\ -0.181 & 0.048 & 0.923*** & -0.431** & -0.269 & -0.468** \end{bmatrix}$$

where *** denotes a parameter significant at the 1% level. ** significant at the 5% level and, * significant at the 10% level. The standard errors are given by:

$$\begin{bmatrix} 0.282 & 0 & 0 & 0 & 0 & 0 \\ 0.248 & 0.167 & 0 & 0 & 0 & 0 \\ 0.254 & 0.144 & 0.137 & 0 & 0 & 0 \\ 0.219 & 0.189 & 0.220 & 0.283 & 0 & 0 \\ 0.223 & 0.237 & 0.299 & 0.323 & 0.409 & 0 \\ 0.147 & 0.130 & 0.121 & 0.172 & 0.195 & 0.184 \end{bmatrix}$$

The element at position (1,1) is the standard error corresponding to the element at position (1,1) in $\hat{L}$. The element at position (2,1) is the standard error corresponding to the element at position (2,1) in $\hat{L}$. Likewise, the standard errors of the remaining elements are given.

B | Programming code

Table 9: *Description of the codes used in this thesis*

Code/file name	Description
haltonDraws	creates Halton matrix for given primes
Numerical experiment	
ML with halton draws with CPP	Optimizes the SLL 100 times with Halton draws and calculates the RMSE.
ML with random draws with CPP	Optimizes the SLL 100 times with random draws and calculates the RMSE.
SLL_ML_random	Function of the SLL calculated using random draws.
SLL_ML	Function of the SLL calculated using Halton draws.
Mixed logit with independent coefficients	
ML with halton draws with CPP	Optimizes the SLL with Halton draws for the true data.
SLL_ML	Function of the SLL calculated using Halton draws.
Mixed logit with dependent coefficients	
ML with MVN	Optimizes the SLL with Halton draws for the true data with dependent coefficients.
SLL_MVN	Function of the SLL calculated using Halton draws with dependent coefficients.
Latent class mixed logit	
LCML with halton draws with CPP	Optimizes the SLL of the latent class mixed logit model with Halton draws .
SLL_cpp	Function of the SLL of the latent class mixed logit model calculated using Halton draws.