ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

BACHELOR THESIS QUANTITATIVE MARKETING AND BUSINESS ANALYTICS

-

INTERNATIONAL BACHELOR ECONOMETRICS AND OPERATIONS RESEARCH

A Comparison of Model Selection Procedures for Latent Class Discrete Choice Models

Author:
Anneleen BROERS (432512)

Supervisor:
MSc. Anoek CASTELEIN

Second Assessor:
Prof. Dr. Patrick GROENEN

Abstract

The use of the latent class model (LCM) is a popular way of modelling discrete choices. It is a good alternative to the traditional multinomial logit model (MNL), but the larger number of parameters comes with a risk of overfitting the data. A good model selection procedure is critical when the LCM is applied in an empirical setting. In this paper, three existing model selection methods are researched and placed in the context of the LCM. The methods are then applied in an empirical setting. Panel data on the purchases of saltine crackers in the Rome (Georgia) market is used to illustrate the methods. This research finds that (i) existing selection methods other than the most commonly used Bayesian Information Criterion (BIC) can provide valuable information during the model selection process and (ii) none of the three methods considered in this paper provide enough information on their own to adequately select a single model.

Keywords Latent Class Analysis - Model Selection

July 4, 2018

1 Introduction

Modelling consumer decisions has been of great interest to both companies and researchers for many years. Understanding consumer behaviour can, for example, help to forecast sales and improve marketing strategies. Many consumer decisions, such as brand choice, can be modelled using discrete choice models. The multinomial logit model (MNL) is a classic example of such a model. However, this model is too restrictive in many application. Hence, researchers have proposed several models that are less restrictive. One popular alternative is the latent class model (LCM) by Greene and Hensher (2003).

This paper compares the LCM to the traditional MNL, where the focus lies on the total number of estimated parameters. As the LCM is an extension of the MNL, it will always give a better in-sample fit. However, this better fit comes at a cost. Because the number of parameters in the LCM is much larger, the estimated model is at risk for overfitting. When overfitting occurs, a model corresponds too much to a particular data set and may fail to fit additional data or create accurate predictions. This research is based around the trade-off that comes with additional parameters in the LCM. Hence, the following research question is formulated:

Which existing model selection methods can be applied to the latent class model by Greene and Hensher (2003) and how do those methods compare?

To answer this question, we first describe the MNL and the LCM in detail. Thereafter, we look at several existing model selection procedures and place those in the context of the LCM. Those procedures are then applied in an empirical setting. Panel data on the purchases of saltine crackers in the Rome (Georgia) market is used to illustrate the different procedures and to select an appropriate model for the data set. Lastly, the selected LCM model is estimated and interpreted.

Previous research has mainly based model selection in a latent class context on information criteria, in particular on the BIC. However, there are many other model selection methods available that can be applied to the LCM. This paper focuses on adapting those methods to fit the LCM and discusses the interpretation of the results.

This research finds that (i) existing selection methods other than the BIC can provide valuable information during the model selection process and (ii) none of the three methods considered in this paper provide enough information to adequately select a single model on their own.

This paper is structured as follows. A brief description of previous research is given in Section 2. We provide descriptions of the MNL and LCM in Section 3 and Section 4 respectively. Different model selection procedures are discussed in Section 5. The empirical panel data used in this paper is described in Section 6. The results are presented in Section 7, followed by a discussion and concluding remarks in Section 8.

2 Theoretical Framework

Modelling discrete choices has been of great interest to researchers over the past decades. A classic and simple model to describe such decision making processes is the multinomial logit model (MNL) by McFadden (1974). Since the development of this model, it has been found to be too restrictive in many applications. The underlying assumptions of the MNL are often unrealistic. One of those important and widely debated assumptions is the independence of irrelevant alternatives (IIA).

Another assumption, that is often found to be too restrictive, is that the fundamental tastes for observed attributes are equal across individuals. This assumption is often unrealistic, especially in the context of consumer decision making. Using a good model for discrete consumer decisions is crucial for developing effective marketing and pricing strategies. Therefore, several extensions of the MNL have been proposed that allow for unobserved taste heterogeneity. Two popular alternatives are the latent class model (LCM) by Greene and Hensher (2003) and the mixed logit model by McFadden and Train (2000).

This paper focusses on the LCM and in particular on the larger total number of parameters in comparison to the traditional MNL. The in-sample fit of a larger model is naturally better, but the additional parameters come with a risk of overfitting the data (Hitchcock and Sober, 2004). When overfitting occurs, a model corresponds to a data set too closely and may fail to fit additional data points. A good model selection procedure is critical to prevent overfitting in a latent class context.

One widely adopted method for model selection in the context of the LCM is the use of model selection information criteria (Lin and Dayton, 1997). In recent literature, especially the BIC value by Schwarz (1978) has been a popular model selection criterion. This method is for example used by Greene and Hensher (2003) and Fiebig et al. (2010). However, there are many more model selection methods available in the literature, applied to a variety of statistical models.

Some alternative methods, discussed by Rao et al. (2001) and others, are hypothesis testing and selection based on the out-of-sample fit of a model. In this paper, hypothesis testing is considered in the form of the likelihood ratio (LR) test. The out-of-sample fit for discrete choice models can be measured by, for example, predictive log likelihood or hit rates. In this paper, hit rates are used as a measure for out-of-sample fit, as they do not penalise incorrect predictions as heavily as the predictive log likelihood measure (Rossi and Allenby, 1993).

3 The Multinomial Logit Model

A traditional model to describe an unordered discrete choice is the multinomial logit model (MNL). The model describes the discrete choice of individual i among J alternatives in T_i choice situations.The probability of choice j by individual i in choice situation t is given by

$$P_{it}(j) = P(Y_{it} = j) = \frac{\exp(x'_{it,j}\beta + z'_{it,j}\gamma_j)}{\sum_{h=1}^{J}\exp(x'_{it,h}\beta + z'_{it,h}\gamma_h)}. \quad (1)$$

The decision by individual i in choice situation t is represented by the random variable Y_{it}. The discrete outcome of that decision is given by y_{it}. In the model, coefficients β correspond to variables $x_{it,j}$ with a generic effect across alternatives. Coefficients γ_j correspond to variables $z_{it,j}$ with an alternative-specific effect.

This model is only identified, if variables $z_{it,j}$ show enough variation over time. To incorporate time-invariant variables, such as a constant, a restriction needs to be imposed on γ_j. Such a restriction on a time-invariant attribute m can be $\gamma_{1m} = 0$. This restriction causes the model to be identified.

In this paper, the MNL is estimated using maximum likelihood estimation (MLE). The choice situations T_i are assumed to be independent. The contribution of individual i, given their choices y_{it}, to the likelihood is

$$P_i(y) = \prod_{t=1}^{T_i}\prod_{j=1}^{J} P_{it}(j)^{I(y_{it}=j)}, \quad (2)$$

where $I(y_{it} = j)$ is the indicator function. This function takes the value 1 if individual i in choice situation t has chosen option j and the value 0 otherwise. Over the entire sample of N individuals, the log likelihood becomes

$$\begin{aligned}\ln L(y;\beta,\gamma) &= \sum_{i=1}^{N} \ln P_i(y) \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{J} I(y_{it} = j)\ln P_{it}(j). \quad (3)\end{aligned}$$

The first order conditions for a maximum can be solved numerically. In this paper, the Berndt-Hall-Hall-Hausman (BHHH) algorithm by Berndt et al. (1974) is used for the maximisation, using the gradient of the log likelihood function in (3). The partial derivatives needed in the BHHH algorithm are given in Section A.1. The MLE is done using the R package mlogit by Croissant (2018).

The standard errors of the parameter estimates can be derived from the hessian matrix of the log likelihood function, evaluated at the estimated parameter vector $\hat{\eta}$. The vector η contains parameters β and γ. The estimated variance of the parameter estimates is given by

$$\hat{V}(\hat{\eta}) = \left(-\frac{\delta^2 \ln L(y;\eta)}{\delta\eta\delta\eta^T}(\hat{\eta})\right)^{-1} \quad (4)$$

The standard errors can be computed by taking the square root of the diagonal elements of the matrix $\hat{V}(\hat{\eta})$.

4 The Latent Class Model

Greene and Hensher (2003) extend the traditional multinomial logit model (MNL) and assume that individuals are implicitly sorted into Q latent classes. The MNL corresponds to $Q = 1$. This latent class model (LCM) for the analysis of discrete choices is based on the assumption that individual behaviour depends on observable characteristics as well as on latent heterogeneity. It is unknown to the analyst which individual belongs to which class.

The model describes the discrete choice of individual i among J alternatives in T_i choice situations. The probability of choice j by individual i in choice situation t, given that the individual belongs to class q, is given by

$$\begin{aligned}P_{it|q}(j) &= P(Y_{it} = j|S_i = q) \\ &= \frac{\exp(x'_{it,j}\beta_q + z'_{it,j}\gamma_{j|q})}{\sum_{h=1}^{J}\exp(x'_{it,h}\beta_q + z'_{it,h}\gamma_{h|q})}, \quad (5)\end{aligned}$$

Random variable S_i represents the class that individual i belongs to. The actual class, which is unknown to the analyst, is given by s_i. To incorporate a time-invariant attribute m, a restriction such as $\gamma_{1m|q} = 0$ needs to be imposed for identification purposes.

The model is estimated using maximum likelihood estimation (MLE). Since the class assignment is unknown, the prior probability for individual i for class q needs to be considered in the likelihood function. In this paper, this probability is assumed to be equal for all individuals and is represented by the parameters θ_q. The restrictions, that apply to these parameters, are

$$\sum_{q=1}^{Q}\theta_q = 1 \quad (6)$$

and

$$0 \leq \theta_q \leq 1. \qquad \text{for } q = 1,\ldots,Q. \quad (7)$$

Choice situations T_i are assumed to be independent. Hence, the log likelihood over the entire sample becomes

$$\ln L(y;\theta,\beta,\gamma) = \sum_{i=1}^{N} \ln \left[\sum_{q=1}^{Q} \theta_q \left(\prod_{j=1}^{J} \prod_{t=1}^{T_i} P_{it|q}(j)^{I(y_{it}=j)} \right) \right]. \quad (8)$$

The log likelihood is maximised using the R package gmnl by Sarrias and Daziano (2017). The Berndt-Hall-Hall-Hausman (BHHH) algorithm is used to directly apply MLE to the log likelihood as given in (8). The package makes use of the analytic derivatives of the log likelihood function in the BHHH algorithm.

To ensure that we find the global maximum using the gmnl package, we use random starting values. The starting values are selected uniformly between -5 and 5. For each model, this random selection is repeated 200 times and each of those times the MLE is carried out for that set of values. Out of the 200 model estimates, the estimate with the highest log likelihood is selected.

The standard errors can be computed in the same way as for the MNL, as described in (4). Parameter vector η contains parameters β, γ and θ.

5 The Model Selection Process

The latent class model (LCM) requires further specification to be applied in an empirical setting. The model incorporates a set of variables, which can either have an alternative-specific or a generic effect on the discrete choice. If a model assumes an alternative-specific effect of the variable, the fit of the model will always be larger than if the model assumes a generic effect. However, the number of parameters increases by $Q \times (J-1)$ for each variable that is assumed to have an alternative-specific rather than a generic effect. Another model selection issue is the choice of Q in the LCM. The parameter Q represents the total number of latent classes. As Q increases, the number of parameters to be estimated goes up rapidly.

Including too many parameters in the model may cause the model to be overfitted. We wish to keep the number of parameters low and the fit of the model high. Several existing methods can help us to decide which variables should have an alternative-specific effect and which value of Q should be chosen. Three of these methods are compared in this paper. The likelihood ratio (LR) test is covered in Section 5.1. The Bayesian Information Criterion (BIC) is described in Section 5.2. The out-of-sample predictive ability is discussed in Section 5.3.

5.1 Likelihood Ratio Tests

To determine whether a variable has an alternative-specific or a generic effect, a likelihood ratio (LR) test can be used. An LR test is generally used to test restrictions on the parameters of a model. It statistically compares the restricted model and the unrestricted model. In the case of the LCM, the restricted model assumes a generic effect of (a) certain variable(s) on choice and the unrestricted model allows for an alternative-specific effect of the variable(s). The LR test statistic (Heij et al., 2004) is given by

$$LR = 2(\ln(L_U) - \ln(L_R)), \quad (9)$$

where L_U corresponds to the likelihood of the unrestricted model and L_R to the likelihood of the restricted model. Under the null hypothesis that the restricted model is correct, it can be shown that

$$LR \xrightarrow{d} \chi^2(g), \quad (10)$$

where g corresponds to the number of parameter restrictions imposed. To test whether a certain variable has an alternative-specific effect on choice, one can look at a series of different models. For every model, there can be an unrestricted model, in which the particular variable has an alternative-specific effect, and a restricted model, in which the variable has a generic effect. When the two models have been estimated, their likelihood can be compared through the LR test to determine whether the alternative-specific effect is significant enough to reject the null hypothesis of a generic effect.

As can be seen from (10), the LR test takes into account the number of parameter restrictions imposed, which corresponds to the difference in the number of parameters between the restricted and the unrestricted model. However, the test does not consider the total number of parameters in either of the models. To capture the effect of the size of the model, a different method needs to be used.

5.2 BIC Values

A measure that captures both the size and the fit of a model is the BIC by Schwarz (1978). This measure gets smaller as the likelihood of the model goes up, but penalises a model with a large amount of parameters. This penalty is incorporated in the form of an added term that increases with the size of the model. The BIC is given by

$$\text{BIC}(p) = -2\ln(L_p) + p\ln(N). \quad (11)$$

This criterion evaluates a model with p parameters applied to a sample of size N. The likelihood of this model is given by L_p. The value of the BIC is small, if the fit of the model is large and the total number of parameters small. Hence, the selected model ideally has a BIC value as small as possible.

This value can be used in the context of determining which variables have an alternative-specific and generic effect, but it does not statistically test for such an effect. The LR test, on the other hand, does test for the significance of an alternative-specific effect. The BIC value is also suitable for giving an insight into the choice of Q, because it takes the size of the entire model into account.

5.3 Out-of-sample Predictive Ability

The ultimate test for overfitting is to compare the out-of-sample predictive performance of different models. To measure this performance, a hold-in sample and a hold-out sample need to be selected from the data set. The model is estimated to fit the hold-in sample and thereafter used to forecast the hold-out sample. Two different methods to forecast the discrete choices in the hold-out sample are used to measure the predictive ability of different models in this paper.

In the context of the LCM, we need to take into account that individuals can belong to any of the Q classes. To get a good estimate of the individual choice probabilities, we can take the expected value over all classes. This can be done using posterior estimates of the latent class probabilities. These estimated probabilities are given by

$$\begin{aligned}\hat{\pi}_{qi|y_i} &= P(S_i = q|y_i) \\ &= \frac{\hat{\theta}_q \prod_{j=1}^{J} \prod_{t=1}^{T_i} \hat{P}_{it|q}(j)^{I(y_{it}=j)}}{\sum_{h=1}^{Q} \hat{\theta}_h \prod_{j=1}^{J} \prod_{t=1}^{T_i} \hat{P}_{it|h}(j)^{I(y_{it}=j)}}. \end{aligned} \quad (12)$$

The expected value of the probability of individual i choosing brand j in an out-of-sample choice situation t^* can be computed as follows:

$$\hat{P}_{it^*}(j) = \sum_{q=1}^{Q} \hat{\pi}_{qi|y_i} \hat{P}_{it^*|q}(j). \quad (13)$$

Using the predicted choice probabilities $\hat{P}_{it^*}(j)$ as a choice probability distribution, two different methods can be used to forecast the discrete choice of individual i in choice situation t^*.

- The Maximum Probability (MP) method: we forecast the discrete choice for which the estimated choice probability is the highest.
- The Monte Carlo (MC) method: we generate a draw from the estimated choice probability distribution and forecast the discrete choice that was randomly drawn.

For the MP method, we can compute the actual out-of-sample hit rate h^*_{MP} as given by

$$\begin{aligned} & h^*_{MP} \\ &= \frac{1}{N^*} \sum_{i=1}^{N^*} \frac{1}{T_i^*} \sum_{t^*=1}^{T_i^*} I\left(y_{it^*} = \underset{j \in \{1,\ldots,J\}}{\arg\max} (\hat{P}_{it^*}(j)) \right), \end{aligned} \quad (14)$$

where N^* is the total number of individuals in the hold-out sample, T_i^* is the total number of choice situations of individual i in the hold-out sample and y_{it^*} is the actual choice of individual i in out-of-sample choice situation t^*.

For the MC method, we can compute a theoretical out-of-sample hit rate h^*_{MC}, that would be observed as the number of randomly drawn forecasts would approach infinity. This theoretical hit rate is given by

$$h^*_{MC} = \frac{1}{N^*} \sum_{i=1}^{N^*} \frac{1}{T_i^*} \sum_{t^*=1}^{T_i^*} \hat{P}_{it^*}(y_{it^*}). \quad (15)$$

6 Data

For the empirical analysis we use optical scanner panel data, collected by the firm Information Research Incorporated. The data set contains information on the purchases of saltine crackers made by 136 households in the Rome (Georgia) market. The data was collected over a two-year period and a total of 3,292 purchases has been recorded. For every purchase observation, the brand choice of the household as well as information on the store and marketing environment at the time of purchase are available. With every visit to the supermarket, the households were presented with several brand alternatives. Among the alternatives, there are three large brands: Nabisco, Keebler and Sunshine with respective market shares of 54.4%, 7.3% and 6.9%. The remaining market share of 31.4% is held by several local brands, grouped together in the data set as "private labels". Each brand alternative has three recorded characteristics, regarding the store and marketing environment at the time of purchase. Those characteristics are referred to in the data set as price, display and feature and are defined as follows:

- Price: the price of a 16-ounce unit of the brand in US dollars at the time of purchase. For the purchased brand, the actual price is used (the shelf price net of the value of any redeemed coupons). For all other brands, the shelf price is used.
- Display: 1 if there was an in-store display of the brand at the time of purchase and 0 otherwise.
- Feature: 1 if there was a newspaper feature advertisement for the brand at the time of purchase and 0 otherwise.

Descriptive statistics of this data set are provided in Table 1. The values for the variables Display and Feature in the table correspond to the fraction of purchase occasions that each brand was on display or featured in a newspaper respectively. The prices in the table correspond to the average sample price of a 16-ounce unit for each brand.

Table 1: Descriptive Statistics for the Purchase of Saltine Crackers (136 households and 3,292 purchases)

	Brands			
Variables	*Sunshine*	*Keebler*	*Nabisco*	*Private Label*
Display[a]	0.129	0.106	0.340	0.099
Feature[b]	0.038	0.043	0.087	0.047
Price[c]($/unit)	0.957	1.126	1.079	0.681
Brand Share	0.073	0.069	0.544	0.314

[a]The display variable corresponds to the fraction of purchase occasions that each brand was on display.
[b]The feature variable corresponds to the fraction of purchase occasions that each brand was featured in a newspaper.
[c]The price variable is the average sample price across all purchase occasions.

7 Results

The data described in Section 6 has been used to illustrate and compare the methods described in this paper. Section 7.1 describes the model selection process in this empirical setting. The results of the likelihood ratio (LR) tests can be found in Section 7.1.1. The BIC values of the different models are considered in Section 7.1.2. The out-of-sample predictive abilities of each model are presented in Section 7.1.3. The selected latent class model (LCM) is estimated and interpreted in Section 7.2.

7.1 Selecting the Model

We wish to select a model, using the selection procedures described in Section 5, that fits the empirical setting described in Section 6. To ensure that the selected model describes the data well and is not overfitted, we compare eight different models based on various statistical tests and metrics. These eight models describe which variables have an alternative-specific effect on choice and which variables a generic effect. The effects of the variables in the different models are specified in Table 2

Table 2: Effects of the Variables for the Models Considered

	Effect	
Models	*Alternative-Specific*	*Generic*
Model 1	None	All
Model 2	Feature	Display and Price
Model 3	Display	Feature and Price
Model 4	Price	Feature and Display
Model 5	Feature and Display	Price
Model 6	Feature and Price	Display
Model 7	Display and Price	Feature
Model 8	All	None

For every model, we consider 1, 2, 3, 4 and 5 underlying latent classes. A model with one latent class is equivalent to a multinomial logit model (MNL).

7.1.1 Likelihood Ratio Tests

LR tests have been performed to test for an alternative-specific effect of each variable. For each variable, two types of LR tests have been performed. In the first set of tests, the restricted model assumes that none of the variables have an alternative-specific effect (Model 1). This restricted model is compared to the unrestricted model, which assumes that only the variable that is being tested has an alternative-specific effect on choice (Models 2, 3 and 4). These tests determine if the alternative-specific effect of the tested variable adds a significant amount of extra value to a small restrictive model.

In the second set of tests. the restricted model assumes that all variables have an alternative-specific effect except for the variable that is being tested (Models 5, 6 and 7). This restricted model is compared to the unrestricted model, which assumes that all variables have an alternative-specific effect on choice (Model 8). These tests determine whether the fit of a large detailed model with many parameters still significantly benefits from incorporating an alternative-specific effect of the tested variable.

For both sets of tests, the maximum log likelihoods of the restricted and the unrestricted model are determined and used to compute the LR-Statistic as in (9) on page 3. The p-value of the test statistic is computed using the asymptotic distribution of the LR-Statistic as in (10). The degrees of freedom for each test is the number of restrictions $g = 3Q$. These LR tests have been executed for models with 1, 2, 3, 4 and 5 latent classes. The results of these LR tests for the variables feature, display and price are presented in Table 3, Table 4 and Table 5 respectively.

Table 3: Results of the Likelihood Ratio Tests for an Alternative-Specific Effect of Feature on Choice

Q	Restricted Log Likelihood Model 1[a]	Unrestricted Log Likelihood Model 2[b]	LR-Statistic	P-Value	Restricted Log Likelihood Model 7[c]	Unrestricted Log Likelihood Model 8[d]	LR-Statistic	P-Value
1	−3347.71	−3339.25	16.92	0.001	−3314.20	−3314.04	0.31	0.959
2	−2328.88	−2317.24	23.27	0.001	−2281.98	−2279.87	4.22	0.646
3	−1959.67	−1952.71	13.91	0.125	−1918.44	−1912.05	12.77	0.173
4	−1832.80	−1822.89	19.82	0.071	−1779.84	−1770.44	18.80	0.093
5	−1756.38	−1745.49	21.79	0.113	−1690.39	−1678.72	23.33	0.077

[a] Alternative-specific: none. Generic: all
[b] Alternative-specific: feature. Generic: display and price.
[c] Alternative-specific: display and price. Generic: feature.
[d] Alternative-specific: all. Generic: none.

Table 4: Results of the Likelihood Ratio Tests for an Alternative-Specific Effect of Display on Choice

Q	Restricted Log Likelihood Model 1[a]	Unrestricted Log Likelihood Model 3[b]	LR-Statistic	P-Value	Restricted Log Likelihood Model 6[c]	Unrestricted Log Likelihood Model 8[d]	LR-Statistic	P-Value
1	−3347.71	−3336.50	22.42	0.000	−3317.11	−3314.04	6.12	0.106
2	−2328.88	−2309.78	38.19	0.000	−2287.79	−2279.87	15.84	0.015
3	−1959.67	−1949.77	19.79	0.019	−1917.93	−1912.05	11.75	0.228
4	−1832.80	−1810.78	44.02	0.000	−1786.89	−1770.44	32.92	0.001
5	−1756.38	−1729.44	53.89	0.000	−1710.93	−1678.72	64.41	0.000

[a] Alternative-specific: none. Generic: all.
[b] Alternative-specific: display. Generic: feature and price.
[c] Alternative-specific: feature and price. Generic: display.
[d] Alternative-specific: all. Generic: none.

Table 5: Results of the Likelihood Ratio Tests for an Alternative-Specific Effect of Price on Choice

Q	Restricted Log Likelihood Model 1[a]	Unrestricted Log Likelihood Model 4[b]	LR-Statistic	P-Value	Restricted Log Likelihood Model 5[c]	Unrestricted Log Likelihood Model 8[d]	LR-Statistic	P-Value
1	−3347.71	−3318.41	58.61	0.000	−3333.82	−3314.04	39.56	0.000
2	−2328.88	−2297.02	63.71	0.000	−2306.65	−2279.87	53.56	0.000
3	−1959.67	−1927.16	65.02	0.000	−1944.13	−1912.05	64.15	0.000
4	−1832.80	−1797.42	70.74	0.000	−1802.89	−1770.44	64.90	0.000
5	−1756.38	−1715.00	82.76	0.000	−1721.58	−1678.72	85.71	0.000

[a] Alternative-specific: none. Generic: all.
[b] Alternative-specific: price. Generic: feature and display.
[c] Alternative-specific: feature and display. Generic: price.
[d] Alternative-specific: all. Generic: none.

From Table 3, Table 4 and Table 5, we can see that the alternative-specific effect of the variable price always turns out to be significant in the LR tests. The results for the tests of the variables feature and display do not show an unambiguous significant alternative-specific effect. Further tests are required to determine whether these variables have an alternative-specific effect or not. Now that we know that the price variable has a significant alternative-specific effect, we can research the null hypothesis that Model 4 is the true model. We test this restricted model against the three less restrictive alternative models. These alternative models are Models 6, 7 and 8. LR tests comparing Model 4 to Models 6, 7 and 8 are given in Table 6, Table 7 and Table 8 respectively.

Table 6: Results of the Likelihood Ratio Tests of Model 4 against Model 6

Q	Restricted Log Likelihood Model 4[a]	Unrestricted Log Likelihood Model 6[b]	LR-Statistic	P-Value
1	−3318.41	−3317.11	2.61	0.456
2	−2297.02	−2287.79	18.46	0.005
3	−1927.16	−1917.93	18.46	0.030
4	−1797.42	−1786.89	21.06	0.049
5	−1715.00	−1710.93	8.15	0.918

[a] Alternative-specific: price. Generic: feature and display.
[b] Alternative-specific: feature and price. Generic: display.

Table 7: Results of the Likelihood Ratio Tests of Model 4 against Model 7

Q	Restricted Log Likelihood Model 4[a]	Unrestricted Log Likelihood Model 7[b]	LR-Statistic	P-Value
1	−3318.41	−3314.20	8.42	0.038
2	−2297.02	−2281.98	30.08	0.000
3	−1927.16	−1918.44	17.44	0.042
4	−1797.42	−1779.84	35.18	0.000
5	−1715.00	−1690.39	49.23	0.000

[a]Alternative-specific: price. Generic: feature and display.
[b]Alternative-specific: display and price. Generic: feature.

Table 8: Results of the Likelihood Ratio Tests of Model 4 against Model 8

Q	Restricted Log Likelihood Model 4[a]	Unrestricted Log Likelihood Model 8[b]	LR-Statistic	P-Value
1	−3318.41	−3314.04	8.73	0.033
2	−2297.02	−2279.87	34.31	0.000
3	−1927.16	−1912.05	30.21	0.000
4	−1797.42	−1770.44	53.98	0.000
5	−1715.00	−1678.72	72.56	0.000

[a]Alternative-specific: price. Generic: feature and display.
[b]Alternative-specific: all. Generic: none.

From Table 6, Table 7 and Table 8, we can see that the null hypothesis is rejected by the alternative models for most values of Q. However, the p-values are in most cases slightly higher than the p-values that originally rejected Model 1 in favour of Model 4. Furthermore, the total number of parameters has not been taken into account in these tests, meaning that the negative effect of the large size of the Models 6, 7 and 8 has not been considered.

To summarise, the LR tests have given us somewhat an idea of the value added by allowing each of the variables to have an alternative-specific effect. However the testing process has been rather lengthy, as a total of 40 models had to be estimated and many tests had to be done. The tests also did not yet provide us with a single optimal model to select.

7.1.2 BIC Values

To gain more insight into the information provided by the different models, we take a look at the BIC value, as described in (11) on page 3. The total number of parameters is given by $p = 7Q - 1$ for Model 1, by $p = 10Q - 1$ for Models 2, 3 and 4, by $p = 13Q - 1$ for Models 5, 6 and 7 and by $p = 16Q - 1$ for Model 8. The BIC value takes the likelihood of the model as well as the total number of parameters into account. According to the BIC, a model is good if the BIC value is small. The BIC values of all eight models are summarised in Table 9.

Table 9: BIC Values of the Models Considered

Q	BIC Values of Model 1[a]	BIC Values of Model 2[b]	BIC Values of Model 3[c]	BIC Values of Model 4[d]	BIC Values of Model 5[e]	BIC Values of Model 6[f]	BIC Values of Model 7[g]	BIC Values of Model 8[h]
1	6744.02	6751.39	6745.90	6709.71	6764.84	6731.40	6725.58	6749.58
2	4763.04	4788.37	4773.44	4747.92	4815.77	4778.06	4766.44	4810.81
3	4081.32	4140.30	4134.42	4089.20	4196.03	4143.63	4144.65	4204.77
4	3884.27	3961.64	3937.44	3910.72	4018.83	3986.85	3972.73	4051.12
5	3788.14	3887.84	3855.74	3826.86	3961.50	3940.20	3899.12	3997.28

[a]Alternative-specific: none. Generic: all.
[b]Alternative-specific: feature. Generic: display and price.
[c]Alternative-specific: display. Generic: feature and price.
[d]Alternative-specific: price. Generic: feature and display.
[e]Alternative-specific: feature and display. Generic: price.
[f]Alternative-specific: feature and price. Generic: display.
[g]Alternative-specific: display and price. Generic: feature.
[h]Alternative-specific: all. Generic: none.

We can see from Table 9 that Model 1 with five latent classes has the lowest BIC value, followed by Model 4 with five latent classes. We can also see that every model has the lowest BIC value for $Q = 5$, meaning that the assumption of five latent classes rather than 1, 2, 3 or 4 seems to be reasonable.

Looking at both the BIC values and the LR tests, one could argue that Model 4 with five latent classes seems to be the best model so far. It assumes an alternative-specific effect of price on choice, which shows an unambiguous significance in the LR tests and it also has the second lowest BIC value out of all considered models.

Both of the methods considered so far have been focussed on preventing overfitting, which occurs when a model fits the data set too closely and fails to accurately predictions in an out-of-sample context. However, these methods have only taken into account the in-sample fit of the different models.

7.1.3 Out-of-Sample Predictive Ability

To examine the out-of-sample predictive ability of the different models, two different prediction methods are considered, as described in Section 5.3. The Maximum Probability (MP) method gives actual out-of-sample hit rates for every considered model. The Monte Carlo (MC) method gives theoretical out-of-sample hit rates.

The models are estimated using a hold-in sample. This sample is made up of the first ten purchase occasions of each of the 136 households. The hold-in sample therefore consists of a total of 1360 purchase occasions. The data of the hold-in sample and the estimated model are used together to determine the posterior estimates of the latent class probabilities. The estimated model and the posterior estimates of the latent class probabilities are then used to predict the choices of the hold-out sample. The hold-out sample consists of the four purchase occasions that directly followed the first ten purchases of each of the 136 households. The total number of purchase occasions in the hold-out sample is 544.

The reason for selecting 10 in-sample and 4 out-of-sample purchase occasions for each household is that every household is treated equally with regards to the estimation procedures. There is an equal amount of information on each household and also an equal amount of predictions to be made. This means that all households are represented equally in the out-of-sample hit rates.

The out-of-sample hit rates of the MP method and the MC method can be found in Table 10 and Table 11 respectively.

From Table 10, we can see that the hit rates of all models are very close to each other. The highest hit rate is the rate of 0.818 for Model 5 with four latent classes. However, because the hit rates in the table are so close together, there is not one single model that consistently performs better than another. It is worth noting that Model 8 (the largest model) predicts less well than smaller models, especially as the number of latent classes increases. This illustrates the risk of overfitting. A large model can create less accurate out-of-sample predictions than a smaller model if it is fitted too closely to the in-sample data points.

From Table 11, we can see that Model 4 with five latent classes has the highest hit rate. A general trend that can be observed from this table is that Models 4, 6 and 7 consistently produce more accurate predictions than the other models. These three models all specify price to have an alternative-specific effect. The hit rate also appears to increase as the number of latent classes goes up to 5. We can again see that the large Model 8 predicts less well than the smaller models.

Using the out-of-sample hit rates, it appears that specifying price to have an alternative-specific effect comes with a better predictive performance. The hit rate also seems to increase with the number of latent classes, at least up to five classes. Considering the LR test results and the BIC values discussed earlier, it seems reasonable to assume that Model 4 with five latent classes is suitable for selection.

Table 10: The Actual Out-of-Sample Hit Rates Using the MP Prediction Method for the Models Considered

Q	*Out-of-Sample Hitrate for Model 1*[a]	*Out-of-Sample Hitrate for Model 2*[b]	*Out-of-Sample Hitrate for Model 3*[c]	*Out-of-Sample Hitrate for Model 4*[d]	*Out-of-Sample Hitrate for Model 5*[e]	*Out-of-Sample Hitrate for Model 6*[f]	*Out-of-Sample Hitrate for Model 7*[g]	*Out-of-Sample Hitrate for Model 8*[h]
1	0.546	0.561	0.577	0.544	0.563	0.583	0.588	0.588
2	0.750	0.744	0.750	0.748	0.748	0.752	0.756	0.752
3	0.796	0.785	0.796	0.785	0.783	0.781	0.787	0.783
4	0.816	0.811	0.807	0.816	0.818	0.779	0.778	0.774
5	0.813	0.798	0.792	0.811	0.790	0.813	0.811	0.765

[a] Alternative-specific: none. Generic: all.
[b] Alternative-specific: feature. Generic: display and price.
[c] Alternative-specific: display. Generic: feature and price.
[d] Alternative-specific: price. Generic: feature and display.
[e] Alternative-specific: feature and display. Generic: price.
[f] Alternative-specific: feature and price. Generic: display.
[g] Alternative-specific: display and price. Generic: feature.
[h] Alternative-specific: all. Generic: none.

Table 11: *The Theoretical Out-of-Sample Hit Rates Using the MC Prediction Method for the Models Considered*

Q	*Out-of-Sample Hitrate for Model 1*[a]	*Out-of-Sample Hitrate for Model 2*[b]	*Out-of-Sample Hitrate for Model 3*[c]	*Out-of-Sample Hitrate for Model 4*[d]	*Out-of-Sample Hitrate for Model 5*[e]	*Out-of-Sample Hitrate for Model 6*[f]	*Out-of-Sample Hitrate for Model 7*[g]	*Out-of-Sample Hitrate for Model 8*[h]
1	0.427	0.427	0.433	0.434	0.429	0.436	0.441	0.437
2	0.635	0.633	0.636	0.656	0.634	0.653	0.655	0.652
3	0.695	0.693	0.696	0.711	0.696	0.707	0.711	0.708
4	0.718	0.719	0.723	0.734	0.721	0.728	0.730	0.727
5	0.739	0.731	0.736	0.755	0.731	0.754	0.751	0.712

[a] Alternative-specific: none. Generic: all.
[b] Alternative-specific: feature. Generic: display and price.
[c] Alternative-specific: display. Generic: feature and price.
[d] Alternative-specific: price. Generic: feature and display.
[e] Alternative-specific: feature and display. Generic: price.
[f] Alternative-specific: feature and price. Generic: display.
[g] Alternative-specific: display and price. Generic: feature.
[h] Alternative-specific: all. Generic: none.

7.2 The Estimated Models

Both the MNL and the selected LCM have been estimated. The results are summarized in Table 12. In the two models, price is specified to have an alternative-specific effect on choice. The display and feature variables are specified to have a generic effect on choice. The model with five latent classes has been selected out of the options 1, 2, 3, 4 and 5 classes.

From the table, we can see that the five classes show different characteristics. We can see that individuals belonging to class 1 are more likely to buy the private label than individuals in class 3. The table also shows the strong alternative-specific effect of price. The choice of an individual in class 1 is for example more negatively affected by a price increase of the brand Sunshine in comparison to a price increase of the brand Nabisco.

We can also see that the standard errors for the classes with a smaller latent class probability (class 3 and 4) are very large for some coefficients. This is something that needs to be taken into consideration when using the LCM. A large amount of uncertainty with regards to the estimated parameters can lead to inaccurate predictions for those smaller classes. One could consider selecting a model with less latent classes or using a sample with a larger number of observations to combat this issue.

Table 12: Estimated Discrete Choice Models (Standard Errors in Parentheses)

Attribute	*Alternative*	*MNL*	*LCM*				
			Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Price[a]	Sunshine	−4.9098***	−9.0735***	−6.0532***	−5.7031***	−2.6895	−6.8738***
		(0.46)	(1.50)	(1.27)	(1.07)	(6.19)	(1.11)
Price	Keebler	−5.7329***	−8.2866*	−6.4392***	−7.7101***	−2.9596	−8.4792***
		(0.67)	(3.44)	(1.67)	(1.85)	(2.07)	(1.95)
Price	Nabisco	−3.2653***	−2.9116**	−1.6726	−8.1214***	−6.9697***	−5.5869***
		(0.28)	(1.06)	(1.00)	(1.20)	(1.81)	(0.77)
Price	Private	−1.4905***	−2.9582**	0.7278	3.0837	−6.6836	0.7467
		(0.26)	(1.03)	(1.28)	(2.42)	(9.60)	(0.69)
Display[b]	All	0.0352	−0.1931	0.2161	−0.4074	0.8494*	0.5733***
		(0.06)	(0.28)	(0.18)	(0.23)	(0.36)	(0.14)
Feature[c]	All	0.4813***	−0.5119	0.9881**	0.5605	0.3767	0.7847***
		(0.10)	(0.46)	(0.33)	(0.33)	(0.90)	(0.20)
Keebler	Keebler	1.7258*	0.4159	1.7628	1.7140	3.9675	2.3713
Constant		(0.81)	(3.85)	(1.93)	(2.14)	(6.63)	(2.24)
Nabisco	Nabisco	1.0096*	−4.0030*	0.6304	3.3981*	6.3738	0.5689
Constant		(0.48)	(1.65)	(1.28)	(1.42)	(6.15)	(1.25)
Private	Private	−2.0481***	−2.1681	−5.4758***	−10.2313***	0.9359	−5.6245***
Constant		(0.44)	(1.33)	(1.24)	(2.16)	(8.29)	(1.09)
Latent Class Probability			0.2477	0.4458	0.0856	0.0479	0.1730
Log Likelihood[d]		−3318.4			−1715.0		

[a]Price is in US dollars per 16-ounce unit.
[b]Display takes the value 1 if there was an in-store display of the brand at the time of purchase and 0 otherwise.
[c]Feature takes the value 1 if there was a newspaper feature advertisement for the brand at the time of purchase and 0 otherwise.
[d]Observations of 136 households making 3,292 purchases.

8 Conclusion

To summarise, this research has discussed the challenges that come with model selection in the context of a latent class model (LCM). As the LCM is an extension of the multinomial logit model (MNL), it will always give a better in-sample fit. As the number of latent classes and the number of incorporated variables gets larger, the in-sample fit increases. However, this better fit comes at a cost. Because the number of parameters in the LCM is much larger, the estimated model is at risk for overfitting. In this paper, several existing model selection methods have been applied to the latent class context and compared in an empirical setting. Panel data on the purchases of saltine crackers in the Rome (Georgia) market was used to illustrate the different procedures.

Likelihood ratio (LR) tests have been discussed and used to test different variables for an alternative-specific effect on choice. These statistical test results have given valuable information about the nature of the incorporated variables. However, the process of testing for all variables is lengthy and the results do not lead to the unambiguous selection of a 'best' model. The LR tests also do not take the total number of model parameters into account. The results are only affected by the number of parameter restrictions imposed.

Another method discussed in this paper is the use of BIC values. These values favour an increased fit of a model, but penalise a model with a large number of parameters. When this value is computed for several models, one can pick the model with the lowest BIC value. This is a very straightforward and unambiguous method, but it could possibly overlook a significant alternative-specific effect of a variable. This was the case for the empirical setting discussed in this paper. The BIC values alone are not powerful enough to accurately select a single model. However, the BIC values are very suitable to be used at the start of the model selection process to create a short-list of a few models to be considered for selection.

The last method discussed in this paper is the evaluation of the out-of-sample predictive ability of different models. Two different ways of forecasting the discrete choices for the hold-out sample have been considered. We have found that the hit rates of different models are often close together and that it can be challenging to draw a conclusion from the comparison of hit rates. However, it was possible to derive certain trends from the comparison. It also became apparent that a model that is too large can actually lead to a lower out-of-sample predictive performance.

In conclusion, we would recommend the following model selection procedure to researchers who are considering using the LCM in an empirical setting.

It is a good idea to first look at a combination of the BIC values and the out-of-sample hit rates to create a short-list of models to consider for selection. LR tests can thereafter be used to statistically test the models against each other.

One of the limitations of this research is that, due to time constraints, only three model selection methods could be compared. There are various other methods available that could be applied to the LCM. Another limitation is that we have only considered models with up to five latent classes in the empirical setting. The estimation of such large models requires much time and computing power, but is worth considering for future research.

A Appendix

A.1 Partial Derivatives of Log Likelihood Functions

The partial derivatives of the log likelihood function of the multinomial logit model (MNL) in (3) on page 2 are given by

$$\frac{\delta}{\delta\beta}\ln L(y;\beta,\gamma) = \sum_{i=1}^{N}\sum_{t=1}^{T_i}\sum_{j=1}^{J} I(y_{it}=j)\left(x'_{it,j} - \frac{\sum_{h=1}^{J_i} x'_{it,h}\exp(x'_{it,h}\beta + z'_{it,h}\gamma_h)}{\sum_{h=1}^{J_i}\exp(x'_{it,h}\beta + z'_{it,h}\gamma_h)}\right) \tag{16}$$

and

$$\frac{\delta}{\delta\gamma_j}\ln L(y;\beta,\gamma) = \sum_{i=1}^{N}\sum_{t=1}^{T_i} z'_{it,j}\left(I(y_{it}=j) - \frac{\exp(x'_{it,j}\beta + z'_{it,j}\gamma_j)}{\sum_{h=1}^{J_i}\exp(x'_{it,h}\beta + z'_{it,h}\gamma_h)}\right). \tag{17}$$

References

Berndt, E., Hall, B., Hall, R., & Hausman, J. (1974). Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement*, *3*(4), 653–665.

Croissant, Y. (2018). Estimation of Multinomial Logit Models in R: The mlogit Packages. Version 0.3-0.

Fiebig, D., Keane, M., Louviere, J., & Wasi, N. (2010). The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity. *Marketing Science, 29*(3), 393-421.

Greene, W., & Hensher, D. (2003). A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit. *Transportation Research Part B: Methodological*, *37*(8), 681-698.

Heij, C., De Boer, P., Franses, P. H., Kloek, T., & Van Dijk, H. K. (2004). *Econometric methods with applications in business and economics.* Oxford; New York: Oxford University Press.

Hitchcock, C., & Sober, E. (2004). Prediction versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1-34.

Lin, T., & Dayton, C. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. *Journal of Educational and Behavioral Statistics, 22*(3), 249-264.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, in *Frontiers in Econometrics*, in P. Zarembka (ed.), New York: Academic Press, 105-42.

McFadden, D., & Train, K. (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics, 15*(5), 447-470.

Rao, C., Wu, Y., Konishi, S., & Mukerjee, R. (2001). On Model Selection. *Lecture Notes-Monograph Series, 38*, 1-64.

Rossi, P., & Allenby, G. (1993). A Bayesian Approach to Estimating Household Parameters. *Journal of Marketing Research*, *30*(2), 171-182.

Sarrias, M., & Daziano, R. (2017). Multinomial Logit Models with Continuous and Discrete Individual Heterogeneity in R: The gmnl Package. *Journal of Statistical Software, 79*(2), 1-46.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464.