MARKETING
ZFP – Journal of Research and Management
www.marketing-zfp.de

Research Articles

Multinomial Logit Models in Marketing – From Fundamentals to State-of-the-Art

By Ossama Elshiewy, Daniel Guhl and Yasemin Boztuğ

Analysing choice behaviour has a long tradition in marketing research. Such an analysis provides valuable insights for researchers interested in understanding consumer behaviour as well as practitioners who aim to optimise their marketing-mix efforts. From this background, our paper gives an overview of the most important aspects when it comes to analysing brand choice using multinomial logit models. Starting with the theoretical foundation of choice behaviour, we move on to summarise the basic models and present the state-of-the-art extensions that account for more realistic choice behaviour. We supplement each model description with an empirical example to emphasise the advantage of each approach compared to the basic models. Finally, we summarise our key findings in the conclusions and highlight avenues for future research. In addition, we provide the estimation code in a web appendix for researchers and practitioners who want to replicate our results or analyse their own research questions using the models described in our paper.

1. Introduction

One of the fundamental elements of marketing research is the analysis of choice behaviour (Russell 2014), whereby researchers aim to identify important determinants that affect decision-makers' choice probabilities. These insights are highly valuable. For instance, they can show how to stimulate brand and product choice in order to enhance brand performance (Shah et al. 2015). They can also strengthen our understanding of psychological and behavioural phenomena, such as reference price formation or loyalty (Fader et al. 1992). Therefore, discrete choice (DC) models have been developed to analyse choices as a function of explanatory variables in a regression-like manner: Decision-makers choose one alternative out of a given choice set (dependent variable), which can be regressed on variables that describe the choice alternatives (explanatory variables). Within the regression framework, DC models hold a multinomial dependent variable (i. e., two or more alternatives) whose outcome can be explained by continuous and/or categorical variables. In a marketing context, the outcome can, for example, be brand choices of different consumers who choose one out of J different brands from one product category (either once or several times). These J brands often differ in terms of their marketing-mix (e. g., price

Ossama Elshiewy is Assistant Professor of Marketing (Akademischer Rat) at University of Goettingen, Platz der Goettinger Sieben 3, 37073 Goettingen, Phone +49 (0)551/39-7269, Fax +49(0)551/39-5849, E-Mail: oelshie@uni-goettingen.de. *Corresponding Author.

Daniel Guhl is Postdoctoral Researcher at Humboldt University Berlin, Spandauer Straße 1, 10178 Berlin, Phone +49(0)30/2093-5750, Fax +49(0)30 2093-5675, E-Mail: daniel.guhl@hu-berlin.de.

Yasemin Boztuğ is Professor of Marketing at University of Goettingen, Platz der Goettinger Sieben 3, 37073 Goettingen, Phone +49(0)551/39-7328, Fax +49(0)551/39-5849, E-Mail: boztug@wiwi.uni-goettingen.de.

Acknowledgements: The authors are thankful to the editor and two anonymous reviewers for their helpful comments during the review process. Furthermore, the authors thank Juliet Reiss for many helpful suggestions regarding the first draft of the paper. Daniel Guhl gratefully acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG) through CRC TRR 190.

or promotion). A DC model will allow estimating the influence of such explanatory variables on the choice behaviour of these consumers. DC models extend models that allow only binary outcomes, such as the logistic regression model or the multinomial logistic regression model, which estimates J–1 binary logistic regressions for multinomial dependent variables with J outcomes.[1]

Historically, DC models can be traced back to Thurstone (1927, reprinted 1959), who laid the theoretical foundation. Luce (1959) extended this groundwork to the concept of choice probabilities, while Tversky (1972) embedded it into the utility framework. Building on these theoretical contributions, McFadden (1973) was the first to formulate a DC model that was in line with the theoretical foundations of choice behaviour and allowed parameter estimates that link the multinomial outcome to explanatory variables. This development led to the well-known multinomial logit (MNL) model that draws from random utility theory to analyse choice behaviour and has become the cornerstone of DC analysis. Professor Daniel McFadden's contribution earned him the Nobel Prize in Economic Sciences [2] in 2000 'for his development of theory and methods for analysing discrete choice'. During the 1970s and 1980s McFadden and his coauthors, as well as other scholars, published numerous papers using MNL models to answer important questions in the field of economics and transportation research (McFadden 2001).[3]

With a certain delay, marketing researchers also began to make use of DC models. Early applications included the pre-test-market evaluation of a newly packaged good (Silk and Urban 1978), the analysis of perceptual mapping data using a MNL model (Hauser and Koppelman 1979), and the analysis of consumers' store choice (Gensch and Recker 1979).[4] Louviere and Woodworth (1983) were the first to introduce Choice-based Conjoint analysis to the marketing discipline, thus laying the foundation for the use of MNL models to analyse such stated-preference data in marketing research. When it comes to revealed-preference data, there is little doubt about the impact of Guadagni and Little (1983), who were the first authors to estimate a MNL model by using consumer purchase sequences collected through supermarket scanner data. Since then, many publications have been devoted to extend the approaches of Louviere and Woodworth (1983) and Guadagni and Little (1983) in order to improve the analysis of choice behaviour in marketing research.

From this background, the goal of this paper is to help academics (Ph.D. students and researchers) and practitioners (e. g., analysts and data scientists) from the field of marketing to learn, understand, and apply MNL models. Therefore, we provide a comprehensive introduction to MNL models, including fundamentals and state-of-the-art extensions, which are supplemented by an empirical example and the corresponding computer code for the complete analysis (data management, estimation, and further steps; see web appendix). We use the statistical programming language `R` because it is freely available and covers a larger number of relevant MNL models compared to commercial software programs (like SAS, SPSS, or Stata).

The structure of this paper is as follows: In Section 2, we provide the theoretical foundations of DC models. We then discuss data management and model specification in Section 3. In Section 4, we present the fundamental MNL models (4.1. covers the basic and Nested MNL), which set the stage for the discussion of the state-of-the-art approaches that have become the gold-standard in marketing research and practice (4.2. examines Latent-class/Finite-mixture MNL, Mixed MNL, and Hierarchical Bayesian MNL). In addition to the discussion, each model is illustrated and compared using the same empirical example. We conclude our paper in Section 5 with a general discussion and avenues for future research.

2. Random Utility Model

The random utility model (RUM) is the theoretical foundation of the MNL models, which are presented in our overview. It was first formulated by McFadden (1973) in the context of DC models and is discussed in detail by Train (2009, p. 14): A decision-maker n faces a choice set with J alternatives ($j = 1,..., J$; with $J \geq 2$). Each alternative j provides a certain level of utility U for decision-maker n that leads to U_{nj}. From the J alternatives, the decision-maker n chooses the alternative i where $U_{ni} > U_{nj}$ $\forall\, j \neq i$ (i. e., the alternative with the highest utility).

The decision-maker n is assumed to know his utility U_{nj} $\forall\, j$. In contrast, the researcher observing the decision-maker does not necessarily possess full information about U_{nj}. Researchers typically only observe choices as well as attributes of the J alternatives (and/or the decision-maker), which can be summarised as the explanatory variables in the vector X_{nj}. A relationship that links X_{nj} to a proxy of the decision-maker's utility is defined as $V_{nj} = f(X_{nj})$. Some aspects of the (true) utility U_{nj} cannot be observed by the researcher, leading to $V_{nj} \neq U_{nj}$. As a consequence, the utility of decision-maker n for alternative j is decomposed into the deterministic (observable) component V_{nj} and the stochastic (unobservable) component ε_{nj}, leading to the well-known RUM:

$$U_{nj} = V_{nj} + \varepsilon_{nj}. \tag{1}$$

ε_{nj} follows a random distribution with density $f(\varepsilon_n)$ and $\varepsilon_n = (\varepsilon_{n1}, \varepsilon_{n2},..., \varepsilon_{nJ})'$. Next, it is possible to specify the choice probability of decision-maker n to choose alternative i, namely P_{ni} as follows:

$$P_{ni} = \text{Prob}(U_{ni} > U_{nj}\ \forall\, j \neq i). \tag{2}$$

Inserting the RUM into this expression yields

$$P_{ni} = \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall\, j \neq i). \tag{3}$$

Rearranging the elements of the inequality then leads to

$$P_{ni} = \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \;\forall\, j \neq i). \quad (4)$$

This latter statement can be interpreted as the cumulative distribution that the difference of the unobserved portion of utility between alternative j and the chosen alternative i ($\varepsilon_{nj} - \varepsilon_{ni}$) is below the difference of the observed portion of utility between these alternatives ($V_{ni} - V_{nj}$). This ensures that the utility is maximised for the chosen alternative in accordance with the observable portion of utility. As soon as the researcher has specified $f(\varepsilon_n)$, this choice probability can be rewritten as

$$P_{ni} = \int I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \;\forall\, j \neq i) f(\varepsilon_n) d\varepsilon_n, \quad (5)$$

where $I(\cdot)$ becomes 1 when the expression in parentheses is true, else 0. The choice probability in Equation 5 is a multidimensional integral over $f(\varepsilon_n)$, and the specification of the density $f(\varepsilon_n)$ defines the particular type of DC model. We will discuss this in detail in the subsequent sections.

The observed portion of utility V_{nj} $\forall\, j$ is specified to describe each alternative j as a function of the explanatory variables, $V_{nj} = f(X_{nj})$. Whereas most applications assume a linear-additive functional relationship, leading to $V_{nj} = f(X_{nj}) = \beta' X_{nj}$, different functional forms for $V_{nj} = f(X_{nj})$ are possible. Typically, researchers observe decision-makers' choices of one alternative out of J alternatives (either once or repeatedly) and then estimate the vector of β parameters (sometimes denoted as utility- or preference-parameters). β determines how the explanatory variables X_{nj} influence the choice probability P_{nj} (respectively utility U_{nj}).

It is important to mention that only differences in utility matter when describing the choice behaviour using the parameters β (i. e., $U_{ni} > U_{nj}$). This has two important implications for parameter estimation and interpretation: First, only explanatory variables that differ across alternatives J can have an impact on choice probability (and utility). For those X_{nj} describing the different alternatives, this is usually unproblematic. These variables either differ across the alternatives (e. g., marketing-mix of different brands) or will not provide relevant information for choice behaviour and are therefore excluded. Explanatory variables that do not differ across alternatives, such as the intercept or any variables that describe the decision-makers, must be constructed in a way that produces the necessary difference across alternatives. This is achieved by defining one alternative r out of the J alternatives as the baseline and then estimating the effect of the intercept (or decision-maker attribute) on the choice probability for each remaining alternative j relative to the baseline alternative r (see, e. g., Train 2009, p. 20).

Second, the scale of utility is irrelevant for the choice behaviour and has to be normalised for identification. In most MNL models this is accomplished by normalising the error terms' variance with a predefined constant (for a detailed discussion on this topic, we refer the reader to Train 2009, p. 23).

3. Data Matrix and Model Specification

This section will provide some considerations and terminology regarding the data matrix and model specification for DC models, which differ from the basic data matrix in regression analysis. Raw data often comes in a different form than the upcoming description and should therefore be transformed before parameter estimation.

A data matrix for a DC model in long format has J rows for each choice situation. If N decision-makers each have T choice situations, each decision-maker has $T{\cdot}J$ rows connected by an ID variable that is repeated $T{\cdot}J$ times. Then the data matrix has in total $N{\cdot}J{\cdot}T$ rows. We will illustrate such a data matrix in the following example, which will be used for all empirical applications in this paper. The data set is from Jain et al. (1994) and has N = 136 consumers that were observed choosing one out of four Cracker brands (Keebler, Nabisco, Sunshine, private-label brand [5]; $j = 1,...,4$, $J = 4$) at several time periods (T between 14 and 77, which constitute different purchase occasions over time). For each consumer n, brand j, and choice situation t, we have the following marketing-mix measures: price per ounce in \$ (price_{njt}), whether there was a newspaper feature advertisement (0/1 dummy: feature_{njt}), and whether there was a display during the choice situation at the point of purchase (0/1 dummy: display_{njt}). In repeated choice situations for fast-moving consumer goods, it is common to include a variable that captures past brand choice behaviour, e. g., the time-varying brand utility that is influenced by past choices (see Ailawadi et al. 1999 for an overview). We use a dummy variable (lastchoice_{njt}) that indicates which brand was purchased at the last purchase occasion (Dubé et al. 2010).[6]

With this example, we can generalise the three indices that are used to describe the variables in DC models: decision-maker n chooses alternative i out of the J alternatives ($j = 1,.., J$) in choice situation t.[7] The alternative-specific variables have indices n, j, and t because they (should) vary across decision-makers, alternatives, and/or choice situations. For parameter estimation, it is important that the information is available for all alternatives J in the choice set and not only for the chosen alternative i. Therefore, in certain cases a data imputation might be appropriate to obtain the necessary data matrix.

In the following example, we show the data matrix in long format for the second to fourth choice situations for Consumer 1 and the last three choice situations for Consumer 136 in the Cracker data set (see *Tab. 1*).

The first variable chid shows the total number of choice situations across all consumers (i. e., in sum 3156). The consumers are identified using id with their consumer-specific choice situation cs. Hence, for the consumer with the id 136, we have 14 choices in the data set. How-

chid	id	cs	choice	alt	price	feature	display	lastchoice
1	1	2	0	keebler	1.09	0	0	0
1	1	2	1	nabisco	.99	0	0	1
1	1	2	0	sunshine	.99	0	0	0
1	1	2	0	private	.71	0	0	0
2	1	3	0	keebler	1.09	0	0	0
2	1	3	0	nabisco	1.09	0	0	1
2	1	3	1	sunshine	.49	0	1	0
2	1	3	0	private	.78	0	0	0
3	1	4	0	keebler	1.09	0	0	0
3	1	4	1	nabisco	.89	0	0	0
3	1	4	0	sunshine	1.03	0	0	1
3	1	4	0	private	.78	0	0	0
				…				
3154	136	12	0	keebler	1.17	0	0	0
3154	136	12	0	nabisco	1.29	0	0	0
3154	136	12	0	sunshine	1.09	0	0	0
3154	136	12	1	private	.59	0	0	1
3155	136	13	0	keebler	1.22	0	0	0
3155	136	13	0	nabisco	1.29	0	0	0
3155	136	13	0	sunshine	1.09	0	0	0
3155	136	13	1	private	.59	0	0	1
3156	136	14	0	keebler	1.04	0	0	0
3156	136	14	0	nabisco	1.23	0	0	0
3156	136	14	0	sunshine	1.29	0	0	0
3156	136	14	1	private	.59	0	0	1

Tab. 1: Data matrix in long format

ever, because we need to initialise the lastchoice variable in the first observation, cs starts with 2 for each consumer. The variable choice describes which of the four brands (alt) were chosen in the particular choice situation. The marketing-mix instruments that applied to the choice situation are summarised in price, feature, and display, while the loyalty measure is shown in lastchoice. We observe that Consumer 1 chooses Nabisco, then switches to Sunshine, and then chooses Nabisco again, while Consumer 136 chose Private in the last three choice situations.

The data matrix in *Tab. 1* allows displaying how the observed portion of utility V is described in terms of the explanatory variables X and the β parameters. We consider the second choice situation (cs = 3, which is the second choice situation because cs = 1 is deleted to initialise lastchoice) by the first consumer (id = 1) to illustrate this relationship (which applies to the entire data matrix). Starting with $V_{njt} = \beta' X_{njt}$, we can write the equations for the observed portion of utility V of Consumer 1 choosing brand j (j = 1,...,4; Keebler = 1, Nabisco = 2, Sunshine = 3, and Private = 4) in choice situation 2 as follows:

$$V_{112} = \beta_1 + \beta_4 \cdot \text{price}_{112} + \beta_5 \cdot \text{feature}_{112} + \beta_6 \cdot \text{display}_{112} + \beta_7 \cdot \text{lastchoice}_{112}$$
$$V_{122} = \beta_2 + \beta_4 \cdot \text{price}_{122} + \beta_5 \cdot \text{feature}_{122} + \beta_6 \cdot \text{display}_{122} + \beta_7 \cdot \text{lastchoice}_{122}$$
$$V_{132} = \beta_3 + \beta_4 \cdot \text{price}_{132} + \beta_5 \cdot \text{feature}_{132} + \beta_6 \cdot \text{display}_{132} + \beta_7 \cdot \text{lastchoice}_{132}$$
$$V_{142} = 0 + \beta_4 \cdot \text{price}_{142} + \beta_5 \cdot \text{feature}_{142} + \beta_6 \cdot \text{display}_{142} + \beta_7 \cdot \text{lastchoice}_{142}$$

Inserting the numeric values for X_{njt} from *Tab. 1* leads to

$$V_{112} = \beta_1 + \beta_4 \cdot 1.09 + \beta_5 \cdot 0 + \beta_6 \cdot 0 + \beta_7 \cdot 0$$
$$V_{122} = \beta_2 + \beta_4 \cdot 1.09 + \beta_5 \cdot 0 + \beta_6 \cdot 0 + \beta_7 \cdot 1$$
$$V_{132} = \beta_3 + \beta_4 \cdot .49 + \beta_5 \cdot 0 + \beta_6 \cdot 1 + \beta_7 \cdot 0$$
$$V_{142} = 0 + \beta_4 \cdot .78 + \beta_5 \cdot 0 + \beta_6 \cdot 0 + \beta_7 \cdot 0$$

In this simplified example, we assume that the parameters β are constant across consumers (see Section 4.2 for extensions to this assumption). We use β_1, β_2, and β_3 as the brand-specific intercepts and fix the intercept for the private-label brand to 0 for identification. The basic idea of these parameter estimates is similar to dummy variable coding for categorical explanatory variables with more than two values: β_1, β_2, and β_3 measure the difference in V of Keebler, Nabisco, and Sunshine compared to the private-label brand. We use one parameter to describe the relationship between the alternative-specific variables price, feature, display, and lastchoice to V, namely β_4 to β_7. These parameters are interpreted globally for all alternatives.[8] For example, we would expect a negative β_4, which means that increasing price will decrease the observed portion of utility (and choice probability) for the consumers.

4. Most Prominent Discrete Choice Models

In this section, we will summarise the most prominent MNL models that have been applied to analyse choice behaviour in marketing research. We will start with the basic models in Section 4.1, which will contribute to the understanding of the more sophisticated approaches in Section 4.2. Throughout these sections, we will emphasise the common foundation of the RUM by extending

Equation 5 to the various upcoming models and briefly discuss the parameter estimation approach. Each model is estimated using the example from Section 3 to illustrate the results.

4.1. Basic Models

4.1.1. MNL Model

The basic MNL model was the first established DC model and was derived by McFadden (1973). Hereby, ε_n in Equation 5 is assumed to follow an i.i.d. Extreme Value (EV) type I distribution. With knowledge about the theoretical density function $f(\varepsilon_n)$ and the cumulative distribution function $F(\varepsilon_n)$, and after some algebraic manipulations (see Train 2009, p. 74), the choice probability in Equation 5 is transformed to the well-known MNL formula. With one choice situation, the choice-probability P for decision-maker n to choose alternative i out of J becomes

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_{j=1}^{J} e^{V_{nj}}} = \frac{e^{\beta' \cdot X_{ni}}}{\sum_{j=1}^{J} e^{\beta' \cdot X_{nj}}}. \tag{6}$$

The numerator is e raised to the power of the observed portion of utility for the chosen alternative i, and the denominator contains the sum of this term for all alternatives J in the choice set. This ratio ensures that the sum of choice probabilities for all alternatives for each decision-maker and choice situation is 1.

The β parameters in Equation 6 can be estimated by maximising the following log-likelihood function: $LL(\beta) = \sum_{n=1}^{N} \sum_{i} Y_{ni} \cdot \ln(P_{ni})$. In this equation, ln transforms the choice probability by the natural logarithm and Y_{ni} is the choice indicator with 1 for the chosen alternative i, else 0 (see choice in *Tab. 1*). The log-likelihood function ensures that the β parameters describe the choice behaviour in a consistent way: For each decision-maker (and choice situation), the predicted choice probability should be highest for the chosen alternative i. As usual for Maximum Likelihood estimation, the Standard Error of β for hypothesis testing is calculated from the square root of the diagonal elements of the inverse of the positive Hessian matrix (see Train 2009, p. 200). To evaluate the overall fit of a MNL model, researchers typically make use of the McFadden R^2 (also called Pseudo R^2) as it allows for interpretations similar to the usual coefficient of determination with values between 0 and 1. The log-likelihood value for the estimated model $LL(\beta)$ is compared to the log-likelihood value of a model with only intercepts, $LL(0)$, and the McFadden R^2 becomes $1 - (LL(\beta) / LL(0))$.

As already mentioned in Section 2, the interpretation of the magnitude of the β parameters is not necessarily meaningful (especially across models). Researchers typically make use of elasticities, as these measures are normalised for the variables' units. The elasticity is the percentage change in one variable that is associated with a one percentage change in another variable. The elasticity of P_{ni} with respect to X_{ni} is calculated using the following equation: $E_{own} = \beta_x \cdot x_{ni} \cdot (1 - P_{ni})$, where β_x is the parameter estimate of the explanatory variable of interest x_{ni}. For the elasticity, the value of P_{ni} comes from the predicted choice probabilities of the model with the estimated β parameters. The cross-elasticity of P_{ni} with respect to x_{nj}, which is an explanatory variable of an alternative other than i, is calculated using $E_{cross} = -\beta_x \cdot x_{nj} \cdot P_{nj}$. The elasticities can be computed for each choice situation and thereafter aggregated across the different values to mitigate aggregation issues (Train 2009, p. 29).

From the cross-elasticities, we can already observe the first limitation of the basic MNL model. It implies proportional substitution across alternatives, referred to as the independence-of-irrelevant-alternatives (IIA) assumption: The ratio of two choice probabilities is independent of the choice probability (and attributes) of other alternatives not included in the ratio (Luce 1959). One consequence is a biased prediction by MNL models, when in reality the ratio of probabilities for two alternatives changes with the introduction or change of another alternative. This leads to the constant cross-elasticities across alternatives in the MNL model. If, for example, one out of the four brands decreases its price, one can assume that the choice probability will increase for this brand (ceteris paribus). As the sum of choice probabilities has to equal 1, the choice probabilities must then decrease for the other three brands. The basic MNL model would imply that the relative decrease in choice probability for these three alternatives is the same, while there are many cases where this does not reflect real choice behaviour. Two further limitations are that the basic MNL model cannot account for unobserved factors in repeated choice situations by the same decision-maker, and that this model assumes common β parameters across decision-makers. These limitations have been successfully solved by the developments that our overview covers in Section 4.2.

Nevertheless, the basic MNL model provides unbiased estimates if the previously mentioned assumptions hold. Two prominent examples from the marketing literature have shown that researchers can obtain valuable insights from basic MNL models by incorporating specific explanatory variables. Guadagni and Little (1983) pioneered adding past choices for modelling brand loyalty. Later, Winer (1986) used the deviations from the individual reference price ('sticker shock') as explanatory variable in addition to the regular observed price to model behavioural price effects.

To illustrate the typical outcome of a MNL model, we analyse the Cracker data set described in Section 3. We estimate a basic MNL model where the choice among the four brands (Keebler, Nabisco, Sunshine, and Private) is explained by the marketing-mix of each brand (price, feature, and display) as well as the last choice of the consumers (lastchoice). The parameter estimates for this model are summarised in *Tab. 2*.

	Estimate	Std. Error
Keebler	**.559**	.143
Nabisco	**1.694**	.125
Sunshine	−.081	.108
Price	**−3.579**	.263
Feature	**.736**	.122
Display	**.175**	.081
Lastchoice	**2.056**	.049

Notes: Statistically significant results ($p < .05$) are indicated in **bold**. The *t*- and *p*-values can be obtained using the R code in the web appendix.

Tab. 2: Parameter estimates for the basic MNL model

	Keebler	Nabisco	Sunshine	Private
Keebler	−3.786	.270	.270	.270
Nabisco	2.061	−1.822	2.061	2.061
Sunshine	.230	.230	−3.214	.230
Private	.745	.745	.745	−1.698

Notes: Elasticities can be interpreted as the percentage change in choice probability for the brand in the column in response to a 1 % change in price for the brand in the row.

Tab. 3: Own- and cross-brand price elasticities for the MNL model

The log-likelihood value for this model is –2100.63 and results in a McFadden R^2 of .378 (because $LL(0)$ is –3375.36), which can be regarded as satisfactory. The brand-intercepts of Keebler and Nabisco are positive and significant, which means that consumers (on average) obtain higher utility from these national brands compared to the private-label brand. The utility of Sunshine does not appear to be significantly different from the latter. However, this 'intrinsic' brand utility excludes loyalty effects (which we capture with the lastchoice variable) and the effect of other covariates (e. g., price and promotion). The β coefficient for price is negative and significant, which means that consumers obtain higher utility from lower prices. Decreasing price will therefore increase choice probability. The parameter estimates for feature advertisement (out-of-store promotion) and display (in-store promotion) are both positive and significant. Hence, using these marketing-mix instruments increases the utility (and choice probability) of the consumers (on average). The β coefficient for lastchoice is also positive and significant. This means that the consumers in the Cracker data set are brand-loyal (on average), with their utility increasing for brands that were chosen in the last choice situation. In contrast, a negative β coefficient for lastchoice would mean that the consumers are variety seeking. However, the magnitude of the effect is most likely overstated because we do not account for heterogeneity in the β coefficients (see Keane 1997; Ailawadi et al. 1999; Dubé et al. 2010).

To compare the elasticities, we calculate the own- and cross-brand price elasticities as described before, which yields the matrix in *Tab. 3*. The diagonal of *Tab. 3* presents the own price elasticities, showing that consumers are more price sensitive for Keebler and Sunshine, while somehow less price sensitive for Nabisco and the private-label brand. The cross-brand price elasticities show changes in the choice probability of the brand in each column by changes in prices for the brand in each row. As already explained, the cross-price elasticities are constant across the alternatives, which means that the basic MNL estimates proportional substitution patterns. In addition, our parameter estimates from *Tab. 2* do not account for unobserved heterogeneity in the β coefficients nor unobserved factors of consumers over time. These assumptions are unrealistic in many cases and can be relaxed through the subsequent extensions of the basic MNL model.

4.1.2. Nested MNL Model

One of the limitations of the basic MNL model is the IIA assumption, which claims that the ratio of two choice probabilities is independent of the attributes of the alternatives not included in the ratio (i. e., proportional substitution pattern). The Nested MNL model mitigates this limitation by separating the choice set into K nests. The IIA assumption then holds only within nests but not across nests. Consider having a choice set of four brands (A, B, C, D). Brands A and C are national brands, while brands B and D are private-label brands. In a Nested MNL model, one can assume that brands A and C are more similar in terms of choice behaviour and that the same holds for brands B and D. The consequence is that a researcher can now predefine nests and assign the alternatives into these labelled nests: Brands A and C are assigned to nest 1, named 'national brands', while brands B and D become nest 2, 'private-label brands'.

In addition to the parameters from the basic MNL model, the Nested MNL model estimates one parameter for each nest k, namely λ_k (k = 1,...,K). These parameters capture the degree of correlation of the alternatives within nest k by $1 - \lambda_k$, with $\lambda_k = 1$ leading to the basic MNL model. Including λ_k into $f(\varepsilon_n)$ and $F(\varepsilon_n)$ of the Generalized EV distribution allows to account for correlations among unobserved factors for the alternatives in the nests. This leads to the following choice probability of decision-maker n for alternative i from nest B_k:

$$P_{ni} = \frac{e^{\frac{V_{ni}}{\lambda_k}} \left(\sum_{j \in B_k} e^{\frac{V_{nj}}{\lambda_k}} \right)^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} e^{\frac{V_{nj}}{\lambda_l}} \right)^{\lambda_l}} \tag{7}$$

The log-likelihood function is the same as for the basic MNL model but with the choice probability from Equation 7: $LL(\beta,\lambda) = \sum_{n=1}^{N} \sum_{i} Y_{ni} \cdot \ln(P_{ni})$.

The Nested MNL model can also be formulated as a product of two basic MNL models (see Train 2009, p. 81), which is typical in marketing and simplifies interpretation: $P_{ni} = P_{nB_k} \cdot P_{ni|B_k}$. The first model ($P_{nB_k}$) captures the decision-makers' choice among the K nests. The second model ($P_{ni|B_k}$) is the choice within the nests, conditional on the choice probability of nest k (given that i is in B_k). For the previously mentioned example, this means that the consumer first chooses between national brands and private-label brands, and then chooses among the brands in the nest. Now, if we decompose the deterministic utility V_{nj} of the decision-maker into W_{nk}, which only varies across nests, and Y_{nj}, which varies within nests, we can express the two probabilities of the Nested MNL model as

$$P_{nB_k} = \frac{e^{W_{nk}+\lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{nl}+\lambda_l I_{nl}}} \text{ and } P_{ni|B_k} = \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}}, \qquad (8)$$

where $I_{nk} = \ln \sum_{j\in B_k} e^{Y_{nj}/\lambda_k}$ is the so-called inclusive value and connects both models. The inclusive value can be interpreted as the expected utility that the decision-maker obtains from the alternatives within the nest B_k.

Many marketing researchers have used this model to analyse whether consumers choose to purchase at all and, if yes, which brand (so-called brand choice-purchase incidence-model; see, e. g., Bucklin and Lattin 1991; Ailawadi and Neslin 1998; Bell and Boztug 2007). Other examples involve using different types of Nested MNL models to show that some consumers first consider brand names and then flavour, while others consider flavour before brand name (Kamakura et al. 1996). The Nested MNL model in our example only mitigates one of the limitations of the basic MNL model. Other researchers in marketing have developed extensions that combine the Nested MNL approach with the concepts from Section 4.2 that relax the other two limitations of the basic MNL model (see, e. g., Ailawadi et al. 2007).

We also attempt to demonstrate an empirical application of the Nested MNL using the data set from Section 3. We define one nest for the national brands Keebler, Nabisco, and Sunshine, while assigning Private to a second nest denoted as 'private-label brand'. This example exhibits one special case of the Nested MNL model: As the second nest only has one alternative, it is important to estimate one common λ for all nests. In the case of two or more alternatives in each nest, it is appropriate to estimate nest-specific λ. The parameter estimates for the latter described model are summarised in *Tab. 4*.

The results of the Nested MNL and the basic MNL model are highly similar with respect to the parameter estimates. Nevertheless, the log-likelihood value increased to –2097.51 and we obtain a McFadden R^2 of .379. The parameter for λ is > 1 but still quite close to one. Usually this coefficient is between 0 and 1 to describe the choice behaviour in terms of utility maximisation. However, Train et al. (1987) argue that this holds if substitution is greater within than among nests, whereas if substitution among nests exceeds substitution within nests, then $\lambda > 1$. Given our nesting pattern, the parameter would be less than one if consumers switch to different national brands more readily than they switch to the different brand types. As λ is (significantly) larger than one, we suppose that consumers switch to different brand types more readily than they switch to different national brands. Hence, the proposed nesting structure does not seem to represent the true substitution pattern among the four brands in our empirical example.

4.2. State-of-the-Art Models

The previous section summarised the basic MNL model. It maintains three assumptions that must be relaxed in specific situations of DC modelling in order to carry out a more realistic analysis. While the Nested MNL model is capable to account for disproportional substitution patterns, most recent advances in MNL models have been concerned with the modelling of heterogeneity in the β coefficients. This means that the researcher assumes different β parameters across decision-makers instead of using one common β parameter to describe choice behaviour. This latter assumption is often unrealistic, although it simplifies the models as well as their estimation. Without heterogeneity, it is not possible to fully understand markets with differentiated products (Allenby and Rossi 1998). Furthermore, models with heterogeneous β parameters fit the (individual choice) data better than their less flexible homogenous counterparts (Ailawadi et al. 1999). Another benefit of including heterogeneity is that such versions of the MNL model allow for

	Estimate	Std. Error
Keebler	**.400**	.157
Nabisco	**1.702**	.125
Sunshine	**−.263**	.129
Price	**−3.695**	.279
Feature	**.791**	.125
Display	**.192**	.079
Lastchoice	**2.174**	.068
λ	**1.136**	.058

Notes: Statistically significant results ($p < .05$) are indicated in **bold**. The t- and p-values can be obtained using the R code in the web appendix.

Tab. 4: Parameter estimates for the Nested MNL model (more) flexible substitution patterns and relax the IIA assumption. Even if one is not interested in heterogeneity itself, ignoring it can lead to biased aggregated results (Abramson et al. 2000). Therefore, this section of our paper will cover the most prominent MNL models that account for unobserved heterogeneity in β parameters.

The common concept of these models is that they define the β coefficients to follow a mixing distribution $f(\beta)$. The choice probability for decision-maker n to choose alternative i out of J is

$$P_{ni} = \int_{-\infty}^{\infty} \left(\frac{e^{\beta_n' \cdot X_{nit}}}{\sum_{j=1}^{MJ} e^{\beta_n' \cdot X_{njt}}} \right) f(\beta) d\beta. \tag{9}$$

However, choice data in marketing applications usually have a panel structure ($T > 1$). The probability for decision-maker n to choose a specific sequence of alternatives $\boldsymbol{i} = \{i_1,..., i_{Tn}\}$ then becomes

$$P_{ni} = \int_{-\infty}^{\infty} \prod_{t=1}^{T_n} \left(\frac{e^{\beta_n' \cdot X_{ni,t}}}{\sum_{j=1}^{J} e^{\beta_n' \cdot X_{njt}}} \right) f(\beta) d\beta. \tag{10}$$

The approach in equation 10 ensures that the β coefficients remain constant within decision-makers (but different across decision-makers).

The fraction inside the parentheses in both equations is the choice probability formula for the basic MNL model from Section 4.1.1, but now the β coefficients have index n to capture the individual-level parameter estimates. The distribution $f(\beta)$ is the so-called mixing distribution. This means that the choice probability becomes a $f(\beta)$-weighted average of the MNL formula at different values of β. The mixing distribution leads to heterogeneous β parameters, depending on how the mixing distribution is defined. This also relieves the ε_n from following only the EV type I distribution but rather having a combination of the latter with the portion of utility that the predefined mixing distribution does not capture in its additional parameters. For example, if $f(\beta) = 1$ for the 'fixed' β and else = 0, the choice probability in Equation 9 becomes the basic MNL model from Section 4.1.1. In the case of a discrete mixing distribution, we have the Latent-class/Finite-mixture MNL model (Section 4.2.1). The Random-parameter/Mixed MNL model uses a continuous mixing distribution (Section 4.2.2). All models can be estimated using either classical or Bayesian inference (see Section 4.2.3). Individual-level β coefficients can be derived irrespective of the statistical approach. While Hierarchical Bayes methods give these estimates as a by-product (Allenby and Rossi 1998), classical approaches employ an additional Bayesian updating step (Revelt and Train 2000). Having individual-level β coefficients is considered the foundation for true one-to-one marketing (i. e., treating each consumer individually instead of applying uniform strategies, see Rossi et al. 1996).

4.2.1. Latent-Class/Finite-Mixture MNL Model

The Latent-class/Finite-mixture (LC) MNL model assumes that $f(\beta)$ is discrete with a finite set of distinct values (β_1, β_2, ..., β_L) (see, e. g., Kamakura and Russell 1989). Researchers have to predefine the number of distinct values L and then estimate a share for each 'latent class' or 'segment', which becomes the mixing distribution $f(\beta)$. In this case the choice probability for decision-maker n to choose alternative i becomes [9]

$$P_{ni} = \sum_{l=1}^{L} S_l \frac{e^{\beta_l' \cdot X_{ni}}}{\sum_{j=1}^{J} e^{\beta_l' \cdot X_{nj}}}. \tag{11}$$

S_l can be interpreted as the share of the population in latent class l and is estimated from the data (Kamakura and Russell 1989). β_l is the latent-class specific parameter vector for the β coefficients. In fact, this procedure estimates L different basic MNL models and each model provides β coefficients that are obtained by weighting the share of the latent class (i. e., a discrete mixing distribution). These different β parameters are considered unobserved heterogeneity in choice behaviour and allow for the estimation of a more nuanced response pattern (i. e., latent-class specific). The IIA assumption holds within the latent classes but is relaxed across latent classes. In repeated choice situations, S_l and β_l remain the same for each decision-maker and therefore capture unobserved factors of decision-maker n over the choice situations.

What makes this approach very appealing is that the heterogeneity distribution can accommodate very flexible unobserved patterns of β parameters if L is chosen to be large enough (Jain et al. 1994). This is also the only model for unobserved heterogeneity that can be estimated by standard Maximum Likelihood because the latent-class share simply enters the log-likelihood function as additional parameters: $LL(\beta,S) = \sum_{n=1}^{N} \sum_{i} Y_{ni} \cdot \ln(P_{ni})$. As this log-likelihood function can have local optima, it is advisable to repeat the optimisation multiple times using different starting values and then compare the solutions. In case the optimisation converges to different solutions, we recommend retaining the result with the highest LL value.

This approach estimates S_l (by the re-parameterisation $S_l = e^{\gamma_l} / \sum_{l'} e^{\gamma_{l'}}$, where we set $\gamma_1 = 0$ for identification) together with the parameter estimates from the L different MNL models. A membership probability for each decision-maker and each latent class can then be derived using conditional Bayes' rule (see Kamakura and Russell 1989). Another approach makes use of the Expectation-Maximization (EM) algorithm that first estimates the membership probability for each decision-maker and each latent class (E-step) and then summarises these to the latent-class share S_l in an iterative procedure (M-step). We refer the reader to Leisch (2004) for a detailed discussion of the EM algorithm for finite-mixture modelling.

	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4	Latent Class 5
ν_l	0*	**1.285**	**.406**	**.442**	**−.303**
(S_l)	(.119)	(.430)	(.179)	(.185)	(.088)
Keebler	**−3.252**	.932	−.502	**1.352**	**7.124**
Nabisco	**−1.991**	**4.197**	**1.813**	**2.342**	**6.349**
Sunshine	**−2.923**	.288	−.303	**.758**	**5.506**
Price	1.477	−1.995	**−11.145**	**−4.793**	**−7.955**
Feature	.053	**1.021**	−.358	**1.031**	.097
Display	.512	.155	.475	**.510**	−.0.146
Lastchoice	**1.024**	**.853**	.394	**.694**	**1.366**

Notes: We omit the Std. Error to save space. Statistically significant results ($p < .05$) are still indicated in **bold**. The ν_l parameter for the first latent class is restricted to 0 for identification and the resulting latent-class shares (S_l) are shown in parentheses. The Std. Errors, *t*- and *p*-values can be obtained using the R code in the web appendix.

Tab. 5: Parameter estimates for the LC MNL model

id	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4	Latent Class 5
7	.000	.001	.000	.046	**.954**
24	**.729**	.000	.271	.000	.000
33	.000	.000	.000	**1.000**	.000
Avg. prob. in assigned latent class	.966	.974	.985	.980	.985
# consumer in assigned latent class	14	62	22	25	13

Notes: The highest membership probability for each consumer is indicated in **bold**.

Tab. 6: Membership probabilities and latent class assignment for the LC MNL model

One can determine the optimal number of latent classes by estimating models with different numbers of L (e. g., two to ten) and then compare these models in terms of model fit (e. g., BIC, see Andrews and Currim 2003 for a discussion of this topic).

The membership probabilities for each decision-maker provide information for a so-called 'fuzzy segmentation' of the decision-makers to the latent classes. To obtain a 'crisp segmentation', each decision-maker is assigned to the latent class with the highest membership probability. The magnitude of the membership probabilities can also measure the quality of the assignment. With this crisp segmentation, the β coefficients of the latent class l describe the choice behaviour of these decision-makers. Estimates of individual-level parameters can be obtained as weighted sums using the membership probabilities as weights.

Several studies have proven the usefulness of the LC MNL model: Kamakura and Russell (1989) showed that several latent classes with different brand preferences and price sensitivities can be revealed by the LC MNL. This approach enabled the authors to analyse competition between national and private-label brands and to investigate complex price-elasticity structures with heterogeneous and disproportional substitution patterns. Similarly, Jain et al. (1994) found disproportional own- and cross-price elasticities for several product categories. Gupta and Chintagunta (1994) proposed a model extension that also incorporates observed heterogeneity (via explanatory variables) to explain the latent-class membership.

To demonstrate the advantage of the LC MNL over the basic MNL, we use the data set from Section 3 to estimate this model extension. To determine the optimal number of latent classes, we estimate LC MNL models ranging from one latent class (which is similar to the basic MNL) to six latent classes. We compare these models using the BIC value. We find that the BIC for five latent classes has the lowest value (3591.89) and therefore present the parameter estimates for this model in *Tab. 5*.

The parameters for the five latent classes show quite some heterogeneity. Consumers in Latent Class 3 are far more price sensitive but less brand-loyal than those in the other latent classes. Consumers in Latent Class 1 are very loyal, but they seem to prefer the private label in general. Regarding brand preferences, the opposite holds true in Latent Class 5 (the smallest segment), where consumers apparently prefer national brands. These data-driven patterns of heterogeneity make intuitive sense. Furthermore, the LC MNL model also fits the data much better than the basic MNL model. We observe a log-likelihood value of –1638.83 and a McFadden R^2 of .514.

Tab. 6 shows the membership probabilities of three randomly selected consumers (id = 7, 24, and 33). We see that the probability values are fairly distinct, making the assignment to latent classes via the crisp segmentation unambiguous and precise. This can also be seen from the

	Keebler	Nabisco	Sunshine	Private
	Latent Class 4			
Keebler	−4.782	.559	.559	.559
Nabisco	2.444	−2.757	2.444	2.444
Sunshine	.569	.569	−4.043	.569
Private	.817	.817	.817	−2.455
	Latent Class 5			
Keebler	−6.270	2.745	2.745	2.745
Nabisco	3.272	−5.360	3.272	3.272
Sunshine	1.822	1.822	−5.833	1.822
Private	.129	.129	.129	−5.302
	Aggregated (over all latent classes)			
Keebler	−4.195	.240	1.042	.141
Nabisco	2.345	−.939	2.254	1.075
Sunshine	.902	.208	−4.087	.347
Private	.320	.289	.791	−1.143

Notes: Elasticities can be interpreted as the percentage change in choice probability for the brand in the column in response to a 1 % change in price for the brand in the row.

Tab. 7: Own- and cross-brand price elasticities from the LC MNL model

average membership probability for assignment of the consumers, which is > .9 in each segment. The last row in *Tab. 6* depicts the number of consumers allocated to each latent class via the crisp segmentation.

To obtain a better understanding of the consequences of accounting for unobserved heterogeneity in LC MNL models, we compare several price elasticity matrices in *Tab. 7*. To conserve space, we only present the elasticity matrices for Latent Classes 4 and 5 as well as the aggregate elasticity matrix obtained by the mixing distribution.[10]

It is clear that the IIA assumption still holds within each latent class but not across latent classes. Consumers in Latent Class 5 have considerably higher elasticities compared to Latent Class 4 and the substitution within national brands is higher. Furthermore, national brands have an impact on the private label, but not vice versa. On average, the aggregated elasticity matrix shows a more nuanced and flexible structure. The IIA assumption no longer holds, because cross-brand elasticities are not the same for the brands (i. e., they are not equal within each row). For instance, the private label has a stronger impact on Sunshine than on Keebler or Nabisco. Keebler on Sunshine have higher cross-elasticities with respect to each other than with respect to Nabisco, but the latter brand sources strongly from both national brands. Compared to the MNL model we can also see that the range of own-price elasticities is larger. The demand of Nabisco is inelastic whereas Keebler and Sunshine now have even higher elasticities (in absolute terms).

In summary, the LC MNL model increases the model fit compared to the basic models, enables the data-driven segmentation of consumers, provides more realistic substitution patterns, and allows the researcher to obtain valuable insights for better marketing-mix decisions.

4.2.2. Random-Parameter/Mixed MNL Model

Instead of a discrete mixing distribution, the random-parameter/mixed (M-)MNL model uses a continuous mixing distribution (e. g., Normal or Uniform distribution). For example, if $f(\beta)$ is assumed to follow a Normal distribution with mean b and covariance W, the density function of β conditional on these parameters will become $\phi(\beta \mid b, W)$ and the general choice probability in Equation 9 will have the following form:

$$P_{ni} = \int_{-\infty}^{\infty} \left(\frac{e^{\beta_n' \cdot X_{ni}}}{\sum_{j=1}^{J} e^{\beta_n' \cdot X_{nj}}} \right) \phi(\beta \mid b, W) d\beta. \quad (12)$$

The aim of the M-MNL model is to estimate the parameters b and W. There is no must to use a full covariance matrix, but restricting the covariance estimates to zero will ignore potential correlations between the individual-level β coefficients. The parameters b and W describe the heterogeneity distribution in terms of their first and second moments. Generally described, the β parameters follow the density function f with parameters θ, which leads to the general mixing distribution $f(\beta \mid \theta)$. The IIA assumption is relaxed because now the ratio of two choice probabilities depends on all attributes of all alternatives in J (see Train 2009, p. 141). Again, unobserved factors of decision-maker n are captured over time by time-invariant β_n.

For parameter estimation[11], it is important to mention that the choice probability in Equation 12 is independent of β_n. These parameters are integrated out, making P_{ni} only a function of θ (e. g., b and W in Equation 12). These choice probabilities have no closed-form solution and therefore a simulation-assisted estimation technique is required to estimate θ. This is usually carried out via a so-called Maximum Simulated Likelihood approach, which employs random draws based on the density $f(\beta \mid \theta)$ and has the following steps: First, draw a value of β from $f(\beta \mid \theta)$ and label this draw β_r with $r = 1$. Second, calculate the choice probability P_{ni} using β_r. Third, repeat the first and second steps R times and average the results for β_r with r = 1,..., R to calculate P^*_{ni}. As a consequence, the simulated log-likelihood function becomes $SLL(\beta | \theta) = \sum_{n=1}^{N} \sum_{i} Y_{ni} \cdot \ln(P^*_{ni})$. Studies by Train (2000) and Bhat (2001) advocate using so-called Halton draws for the first step instead of (pseudo-)random draws to reduce simulation noise and increase the accuracy of estimating $f(\beta \mid \theta)$. In general, using many draws will reduce variation in LL values due to simulation noise ($R >$ 1000 or $R >$ 5000 is recommended in the case of many explanatory variables; see, e. g., Elshiewy et al. 2017), thereby improving convergence to the global maximum and minimising the risk of masking empirical non-identification (Choiu and Walker 2007). While the problem of starting values is less pronounced for continuous mixing distributions, we recommend checking multiple starting values, as in the case for LC MNL models.

Most applications assume that $f(\beta \mid \theta)$ follows a Normal distribution (Chintagunta et al. 2005), but some researchers also make use of Lognormal distributions to force the values of $f(\beta \mid \theta)$ to have the same sign (e. g., for the β parameter of price; see Kim et al. 1995). However, the distribution of $f(\beta \mid \theta)$ can also follow a Uniform, Truncated Normal, Gamma, or other distribution (see, e. g., Henscher and Greene 2003). In a recent publication, Train (2016) developed an approach to use highly flexible distributions for $f(\beta \mid \theta)$. With this technique, the researcher has to define variables to describe the mixing distribution, which results in flexible forms such as polynomials, splines, or step functions for the heterogeneity distribution. These approaches have not yet been transferred to the marketing discipline and could be a promising avenue for future research.

It is important to mention that the Maximum Simulated Likelihood approach does not provide an assignment of the individual-level β to the decision-makers but only estimates $f(\beta \mid \theta)$ such that one can describe the heterogeneity distribution by θ. Revelt and Train (2000) demonstrate how the assignment can be conducted by using Bayes' rule and simulations of conditional expectations of β. However, the Hierarchical Bayesian counterpart of the M-MNL model (see Section 4.2.3) provides the desired assignment with parameter estimation and should be regarded as more 'straightforward' for obtaining individual-level parameters.

In marketing, the unobserved heterogeneity in β is interpreted as heterogeneity in consumer sensitivity to marketing instruments and as variation in brand preferences captured by the intercepts or the loyalty measure. This is a highly important issue that has led to numerous publications in marketing research making use of the M-MNL model. For example, Gönül and Srinivasan (1993) established that consumers react differently to marketing actions and have heterogeneous intrinsic brand preferences. Therefore, their (aggregated) price elasticity matrices do not exhibit the typical (unrealistic) proportional patterns. Erdem et al. (2008) analysed multiple fast-moving consumer goods categories (i. e., toothpaste, toothbrushes, ketchup, detergent) and found that for most brands, TV advertising increases consumers' marginal willingness-to-pay for a brand. They establish that accounting for consumer heterogeneity is crucial for a solid understanding of advertising effects. Horsky et al. (2006) used a M-MNL model to analyse toothpaste brand choice and added observed heterogeneity in the form of individual-level brand preferences ('brand liking') from survey data of the same consumers.

Again, we will illustrate the M-MNL model as extension to the basic MNL model using the data set from Section 3. We estimate a M-MNL with normally distributed random parameters for both the intercepts and the explanatory variables. We use 1000 Halton draws to simulate the log-likelihood function. In addition to the heterogeneity distribution, the covariance W is estimated to allow for correlation patterns across the β coefficients. The mean parameter estimates as well as their standard deviations (a measure for the heterogeneity; see Chintagunta et al. 2005) are summarised in *Tab. 8*.

There is some difference between the magnitudes of the average β of the random parameters from the M-MNL model and the parameter estimates from the basic MNL. Most noteworthy is that the parameter estimate for the (mean) display effect has become insignificant and that the (mean) effect of brand loyalty (lastchoice) is now less than one-quarter of the estimate in the MNL model. The latter result is in line with previous research: as soon as heterogeneity is accounted for, the effect of brand loyalty remains significant but becomes smaller (see Keane 1997 and Dubé et al. 2009). Furthermore, all standard deviations are significant and their magnitudes reveal a

	Mean		Standard Deviation	
	Estimate	Std. Error	Estimate	Std. Error
Keebler	.272	.377	**5.328**	.561
Nabisco	**3.350**	.358	**5.274**	.497
Sunshine	.220	.351	**4.111**	.412
Price	**−3.711**	.651	**5.669**	.692
Feature	**.831**	.256	**1.132**	.285
Display	.228	.159	**.960**	.195
Lastchoice	**.447**	.120	**.354**	.139

Notes: Statistically significant results ($p < .05$) are indicated in **bold**. The t- and p-values can be obtained using the `R` code in the web appendix.

Tab. 8: Parameter estimates for the M-MNL model

	Keebler	Nabisco	Sunshine	Price	Feature	Display	Lastchoice
Keebler	1						
Nabisco	.833	1					
Sunshine	.952	.864	1				
Price	−.468	−.144	−.563	1			
Feature	−.140	−.018	−.203	.365	1		
Display	−.259	−.476	−.341	−.330	−.356	1	
Lastchoice	.327	.265	.172	.306	−.028	−.100	1

Notes: We omit the upper triangle of the correlation matrix to facilitate readability.

Tab. 9: Correlation of β coefficients for the M-MNL model

	Keebler	Nabisco	Sunshine	Private
Keebler	−3.665	.203	1.131	.175
Nabisco	1.700	−.711	2.568	.394
Sunshine	.858	.233	−5.406	.366
Private	.489	.137	1.377	−.642

Notes: Elasticities can be interpreted as the percentage change in choice probability for the brand in the column in response to a 1 % change in price for the brand in the row.

Tab. 10: Own- and cross-brand price elasticities for the M-MNL model

considerable amount of unobserved heterogeneity in all β parameters. For example, the values of the standard deviations imply that some consumers have a negative effect of lastchoice (i. e., variety seeking), while others have low or even positive price sensitivities or are not necessarily positively influenced by promotions. The intrinsic brand preferences are very heterogeneous as well. The log-likelihood value of the M-MNL model is –1583.75 and results in a McFadden R^2 of .531. Hence, this model fits the data better than the LC MNL model with five latent classes, even though it has fewer parameters (35 vs. 39).

Tab. 9 summarises the correlation pattern of the β coefficients for the intercepts and the explanatory variables. The correlation pattern shows very interesting findings for choice behaviour. We observe for some consumers that a display does not increase choice probability of the brands (see *Tab. 8*). One possible explanation can be found in the correlation pattern in *Tab. 9*. The negative correlation between the random parameters of lastchoice and display (–.100) reveals that more loyal consumers are less display sensitive. In addition, more loyal consumers are (ceteris paribus) less price sensitive because of the positive correlation (.306). The positive correlation between all national brands is also intuitive. These findings are highly valuable for a deeper understanding of brand choice behaviour and allow customising the marketing-mix efforts to consumers with greater sensitivity in choice probability.

Lastly, we look at the price elasticities from the M-MNL model to gain a better understanding of the substitution patterns in the Cracker category. Because of the mixing distribution, the elasticities have to be calculated via simulation. The price elasticities in *Tab. 10* are more or less similar to the aggregated price elasticities from the LC MNL model (*Tab. 7*). However, the demand for the private-label brand is now inelastic on average (E_{own} < 1) and the elasticity for Sunshine is now even higher.

4.2.3. Hierarchical Bayesian MNL Model

Bayesian estimation has gained increasing popularity in marketing (Rossi and Allenby 2003).[12] Here the classical likelihood principle is combined with prior beliefs for model parameters, denoted as prior distribution. Bayesian inference is carried out by summarising the so-called posterior distribution, which is proportional to the likelihood times the prior distribution. Numerous advances in Markov Chain Monte Carlo (MCMC) methods have led to Hierarchical Bayesian (HB) approaches for estimating MNL models with heterogeneity (see, e. g., Rossi and Allenby 1993; Allenby et al. 1995; Allenby and Ginter 1995). HB approaches employ first- and second-stage priors, which lead to the term 'hierarchical'. Here, the first-stage prior is the mixing distribution in Equation 9, which determines the individual-level parameters β_n. The second-stage prior includes the prior beliefs regarding the parameters that define the first-stage prior (i. e., b and W in the case of a Normal mixing distribution as in Equation 12).

The parameters (β_n, b, and W) are not obtained via optimisation but by repeatedly sampling draws from the posterior distribution. This results in a sequence (Markov chain) for each parameter that can be summarised after convergence has been achieved. Convergence is hereby defined as reaching a stationary distribution for a reasonable number of draws. The number of draws needed for reaching convergences (discarded as burn-in) as well as the number of draws necessary for precisely summarising the posterior distribution (after burn-in) are chosen by the researcher and should be supported via formal tests or visual inspection (see Gelman et al. 2015 for a detailed discussion).

One major advantage of the HB MNL model is that it directly provides the assignment of the individual-level β coefficients to the decision-makers by summarising the MCMC draws for each decision-maker (while the classical approach requires an additional Bayesian updating step; see Revelt and Train 2000). Early applications of

HB MNL models have assumed a Normal distribution for the heterogeneity (Allenby et al. 1995; Allenby and Ginter 1995), which has subsequently been carried out by numerous studies in the marketing literature (for a review, see Rossi and Allenby 2003). Studies have shown that using a Normal distribution as first-stage prior leads to similar results as using a Normal distribution for the previously described M-MNL model (see Elshiewy et al. 2017; Huber and Train 2001). However, there is increasing criticism regarding the Normal distribution for consumer heterogeneity, as it is unable to represent a potential multi-modal heterogeneity distribution (Gilbride and Lenk 2010). The advantage of the HB MNL model is that it allows for a convenient and robust estimation of more flexible heterogeneity distributions for β_n. Early attempts by Kalyanam (1996) as well as Allenby et al. (1998) proposed a so-called mixture of Normals prior, which allows for the estimation of skewed and multi-modal heterogeneity distributions. This approach assumes that the first-stage prior follows C Normal distributions, where individuals belong to one of the C Normal distributions, similar to the crisp segmentation in the LC MNL model (Section 4.2.1.). The mixture of Normals prior (denoted as a semi-parametric approach) can be exchanged to even more flexible heterogeneity distributions (non-parametric distributions). For a detailed treatment of HB MNL models with semi- and non-parametric heterogeneity distributions, the reader is referred to Rossi (2014a).

Some examples from marketing research have highlighted the strengths of HB MNL models. Ainslie and Rossi (1998) estimated a multi-category choice model in which the household response coefficients were assumed to be dependent across categories. This estimated distribution of heterogeneity revealed that price, display, and feature sensitivity are not uniquely determined for each category and may be related to household-specific factors. Dubé et al. (2008) demonstrated that brand loyalty should have a strong effect on a retailer's category pricing decision. To analyse loyalty-effects in DC models, heterogeneity in β coefficients must be accounted for because otherwise, these effects can be upward biased, leading to wrong implications (Keane 1997). In a companion article, Dubé et al. (2010) further analysed this issue and found strong evidence that observed inertia in brand choices is robust to preference heterogeneity by using a mixture of Normals prior distribution for the β coefficients. In line with these developments, Gilbride and Lenk (2010) proposed an approach to determine whether the Normal prior distribution can be considered an adequate choice for describing the heterogeneity of decision-makers or if more flexible priors should be favoured.

For our empirical example using the Cracker data set, we also estimate a HB MNL model with a Normal prior using 200,000 draws. Convergence of the MCMC chains is observed after 100,000 draws (burn-in). We keep every 100th draw (thinning) after burn-in and therefore remain with 1000 draws for summarising the posterior distribution. As usual in Bayesian inference, we present the posterior means of the MCMC draws as parameter estimates for each explanatory variable and summarise the uncertainty by the 95 % credibility interval (CI) of these distributions (Rossi et al. 2005). These estimates are summarised in *Tab. 11*. These estimates are very similar to the results from the M-MNL model, which is in line with previous research (Elshiewy et al. 2017, Huber and Train 2001).[13]

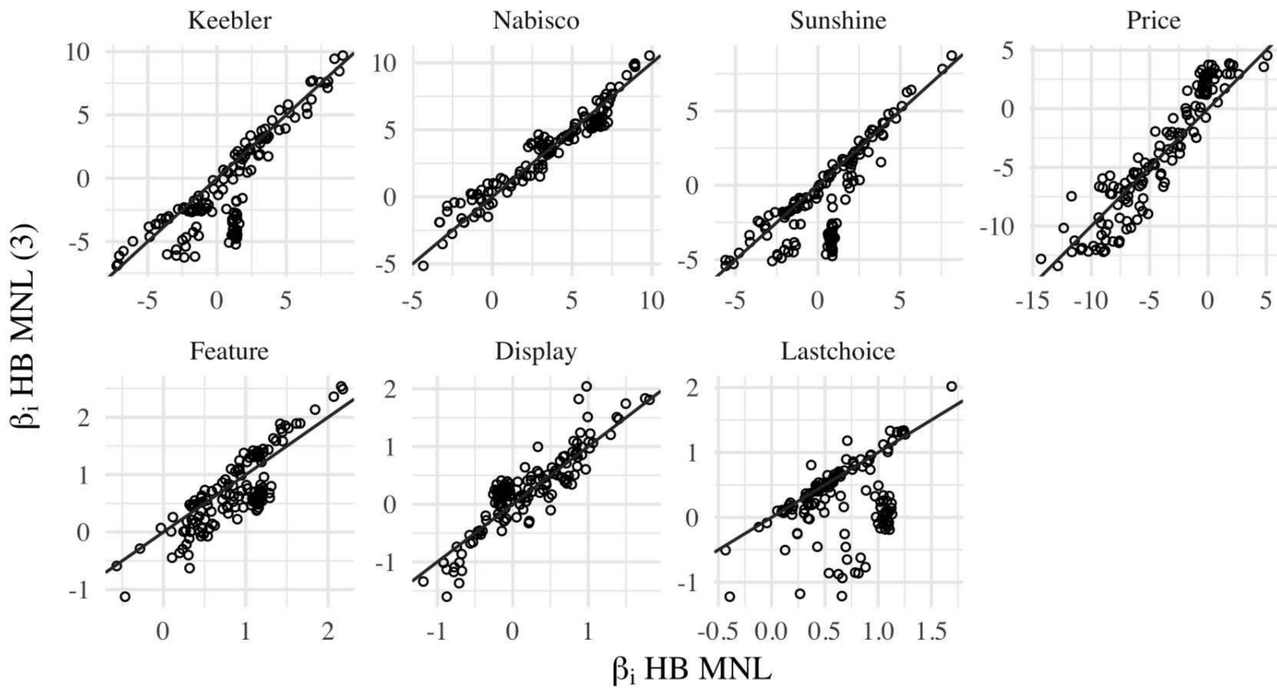
More interesting is the analysis of the HB MNL model that uses the mixture of Normals prior instead. We estimated models with one, two, three, five, and ten components and picked the model with three components because it has the highest trimmed log marginal density (LMD) value of –1236.08 (Dubé et al. 2014). Plotting the LMD against the number of components typically shows an 'elbow'-like pattern; after a certain number of components, the LMD does not increase anymore or even decreases (because LMD automatically accounts for model complexity, see Rossi et al. 2005). We recommend using the number of components at the elbow because it represents a good compromise between flexibility and parsimony.

We employ the same number of draws as before, but we do not summarise the moments of the posterior marginal distributions of the HB MNL model with three components, as they are comparable to the ones reported before for the HB MNL model with the Normal prior (*Tab. 11*). Instead, a comparison of the individual-level β coefficients is insightful. *Fig. 1* presents scatterplots for each parameter in both HB MNL models.

	Mean		Standard Deviation	
	Posterior Mean	95 % CI	Posterior Mean	95 % CI
Keebler	.865	(−.183, 1.958)	**4.239**	(3.341, 5.307)
Nabisco	**3.636**	(2.733, 4.676)	**3.981**	(2.997, 5.115)
Sunshine	.409	(−.396, 1.297)	**3.199**	(2.499, 4.006)
Price	**−4.125**	(−5.590, −2.692)	**5.245**	(3.680, 7.004)
Feature	**.859**	(.411, 1.308)	**1.098**	(.769, 1.511)
Display	.203	(−.152, .541)	**1.040**	(.794, 1.370)
Lastchoice	**.663**	(.364, .937)	**.763**	(.587, 1.012)

Notes: Estimates in **bold** indicate that the 95 % credibility interval does not cover zero.

Tab. 11: Parameter estimates for the HB MNL model



Notes: The scatterplots depict estimated individual-level β coefficients for the HB MNL model (x-axis) and the HB MNL model with three components (y-axis). Deviations from the 45° line indicate differences in the estimates across both models. Each sub-plot corresponds to a specific parameter (e. g., brand-specific constant of Keebler or price sensitivity).

Fig. 1: Individual-level β coefficients from the HB MNL model (1 vs. 3 components)

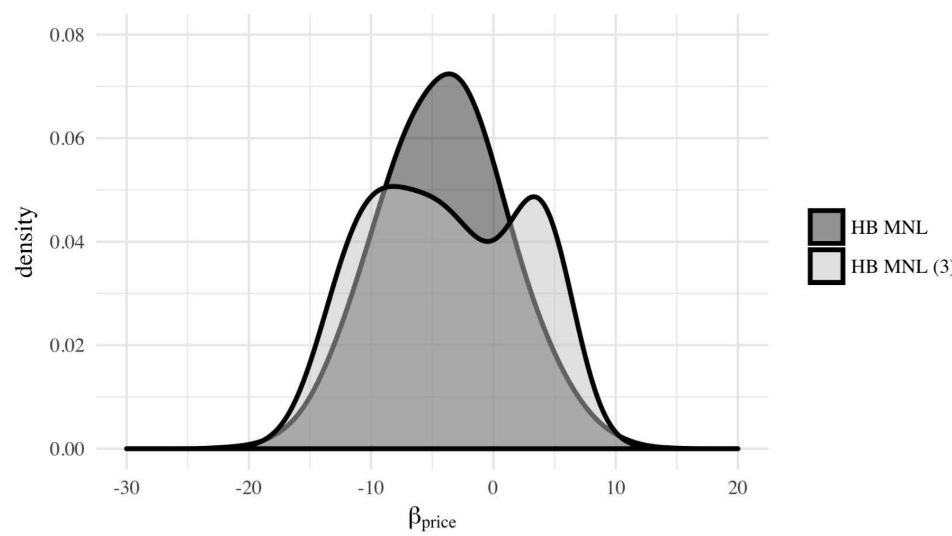


Fig. 2: Marginal distributions of the β parameter of price (HB MNL model, 1 vs. 3 components)

The non-normality of the prior with three components has a clear influence on the results of the individual-level β coefficients, indicated by deviation from the 45° line. Some estimates tend to be more extreme (e. g., intercept for Keebler, price, or the lastchoice effect), which is due to the greater flexibility of the mixture of Normals prior.

Fig. 2 shows the marginal distribution of the β parameter of price from the HB MNL and HB MNL (3) model in a density plot. Both models show a parameter distribution with values of similar magnitude. Most of the probability mass is in the negative domain, and the average values as well as the range of the distributions are comparable. However, the HB MNL model with three components clearly identifies a multimodal distribution, which the HB MNL with the regular Normal distribution cannot entertain.

5. Conclusion

There is little doubt about the importance of analysing individuals' choice behaviour. Numerous scientific publications and practical applications in marketing research have proven the benefits of analysing and predicting brand choice. Such results provide valuable insights for brand managers when it comes to optimising marketing-mix efforts (Shah et al. 2015) and furthermore have been successfully applied to study behavioural aspects of consumer behaviour (Fader et al. 1992). From this background, our paper provided an overview of the most relevant models that are common in marketing research and practice. While our empirical example made use of revealed-preference data, all aspects of parameter estimation and model interpretation are similarly applicable to stated-preference data (e. g., Choice-based Conjoint data).

Starting with the basic MNL model, we have demonstrated how to estimate, test, and interpret the β coefficients of the explanatory variables that are assumed to influence the (brand) choice behaviour. We also advocate the use of elasticities to account for the dimension of the explanatory variables and compare the effects in terms of percentage changes. It is important to keep the three assumptions in mind that hold for the basic MNL and can be considered as limitations in particular situations. These limitations are (i) proportional substitution patterns (IIA assumption), (ii) no random taste variation due to common β coefficients across decision-makers, and (iii) no consideration of unobserved factors over time in repeated choice situations.

The Nested MNL model can account for disproportional substitution patterns by classifying the choice alternatives into predefined subgroups (nests). This model provides additional parameter estimates compared to the basic MNL model to estimate the substitution pattern within the nests. Our empirical application offers a good example in which such a predefined nesting structure (that theoretically makes sense) does not provide the best representation of the substitution pattern that exists in the brand choice behaviour of the consumers.

To address the three limitations of the basic MNL model, our paper summarised the state-of-the-art MNL models that relax all three limitations and account for more realistic choice behaviour. The common approach of these extensions is to allow the β coefficients to vary across decision-makers. In a brand choice setting, this variation is interpreted as consumer heterogeneity in sensitivity to changes in the marketing-mix instruments (or sensitivity in behavioural drivers, such as loyalty). As soon as such random taste variation is included in the MNL model, the other two limitations can be relaxed. First, disproportional substitution patterns can be accounted for, as the ratio of two alternatives (as well as the cross-alternative elasticities) now depends on changes in all alternatives in the choice set (IIA assumption relaxed). Second, the individual-level β coefficients remain constant over time and can therefore account for unobserved factors of decision-makers in repeated choice situations. For our empirical example, we estimated the LC MNL with five distinct β coefficients per explanatory variable, the M-MNL and HB MNL with normally distributed β coefficients, and the HB MNL with a three-component mixture of Normals heterogeneity distribution for β. We showed how accounting for heterogeneous β coefficients affects the (average) parameter estimates compared to the basic MNL, and how the knowledge of heterogeneous marketing-mix sensitivities across consumers provides valuable insights for understanding, predicting, and influencing brand choice behaviour. From this, we emphasise the importance of using the state-of-the-art approaches, even if only interested in aggregated parameter estimates. This will avoid biased inference by accounting for the most realistic choice behaviour. While the more flexible heterogeneity distributions, like HB MNL with mixture of Normal prior, can be considered to account for the more realistic choice behaviour compared to the models using parametric heterogeneity distributions, we recommend using these models with care, as greater flexibility often comes with more demand with respect to the data. To replicate our results and to offer the opportunity to apply the models in our paper to own research questions, we provide the `R` code for data management and parameter estimation in the web appendix.

While discrete choice modelling has a long history in marketing (Russell 2014), we believe that there are still various opportunities for future research. This paper focused on MNL models (with and without preference heterogeneity) and covered classical and Bayesian estimation for data on the individual level. However, starting with applications in Empirical Industrial Organisation (see Berry 1994; Berry et al. 1995), there is a growing interest in marketing research on estimating DC models with preference heterogeneity using aggregate data (i. e., market shares of J alternatives as the dependent variable). Employing Generalized Method of Moments (Nevo 2000), Maximum Likelihood (Park and Gupta 2009), or Bayesian estimation (Jiang et al. 2009; Zenetti and Otter 2014), these models typically also account for (price) endogeneity which can be an issue in the case of revealed-preference data, and is an often and controversially discussed topic in marketing research (Villas-Boas and Winer 1999; Shugan 2004; Chintagunta et al. 2005; Petrin and Train 2010; Rossi 2014b). Another interesting topic are MNL models that account for so-called scale heterogeneity (i. e., individual-specific error term variances). Fiebig et al. (2010) proposed the Generalized MNL (G-MNL) model, which is supposed to explicitly capture scale heterogeneity in addition to preference heterogeneity. However, some scholars are sceptical whether it is possible to empirically disentangle scale and (correlations in) preference heterogeneity (Hess and Rose 2012; Hess and Train 2017). As the capabilities and limitations of the G-MNL model are not yet fully understood, more research on this topic will be helpful. Furthermore, research has also investigated models assuming that decision-makers use individual-specific subsets of alternatives ('choice set heterogeneity', see Bronnenberg and Vanhonacker 1996; Van Nierop et al. 2010) or subsets of attributes ('attribute non-attendance', see Hensher et al. 2005; Hole 2011; Yegoryan et al. 2016). These aspects are valuable additions to the toolbox of applied choice modellers because assuming full information processing by consumers with respect to all alternatives and attributes in a choice set can be unrealistic in some cases (see Currim et al. 2015). Lastly, due to the ever-increasing size of consumer databases, the applicability and scalability of models and estimation methods are of crucial importance and should be explicitly considered in future research (Frischknecht et al. 2014; Braun and Damien 2016; Chintagunta et al. 2016; Bradlow et al. 2017). We expect that recent advances in machine learning will have an impact on DC modelling and marketing research in general (Jacobs et al. 2016; Sudhir 2016).

Notes

[1] Readers interested in these models are recommended to consult Agresti (2013).

[2] Correctly expressed: Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel.

[3] Hausman and Wise (1978) developed the multinomial probit (MNP) model as an alternative to the MNL model for analysing choice behaviour but with different distributional assumptions. The state-of-the-art extensions in DC modelling (see Section 4.2) can be accommodated in a more convenient way using the MNL model compared to the MNP model (i. e., the GHK sampler for calculating MNP choice probabilities is cumbersome to program and software for heterogeneous MNL models is nowadays widely available). Hence, the MNL model has become more prominent in marketing research and practice. Therefore, our paper focuses on MNL models. Readers interested in the MNP model are referred to Train (2009, Ch. 5).

[4] Before the era of DC models in marketing, so-called stochastic brand choice models were developed. We refer the interested reader to Montgomery et al. (1970) for an overview of this topic.

[5] Hereafter also denoted as 'Private'.

[6] Another common past choice measure is the exponentially smoothed loyalty variable of Guadagni and Little (1983). Both versions capture brand loyalty and typically improve the fit of brand choice models considerably. However, to initialise the lastchoice variable we only need to exclude one observation for each household, estimating an additional smoothing parameter is avoided (e. g., via grid-search), and derived market segments tend to be more stable (Ailawadi et al. 1999).

[7] For data with only one choice situation per decision-maker, the index t is irrelevant.

[8] While this approach is common, it is also possible to estimate one β for each alternative to describe the relationship between the alternative-specific variables and V (see, e. g., Krishnamurthi and Raj 1988).

[9] For simplicity we omit the index t for the choice situations. In the case of multiple choice situations, the weighted sum in Equation 11 is calculated over the sequence of choices for each n (see Equation 10).

[10] For computing elasticities in the LC MNL model, one can either use the formulas in Kamakura and Russell (1989) for weighting the segment-specific elasticity matrices or use simulation.

[11] In this section, we consider the classical parameter estimation approach, while the Hierarchical Bayesian counterpart is covered in Section 4.2.3.

[12] We do not discuss Bayesian inference and MCMC for Hierarchical Bayesian models in detail. Interested readers from the field of marketing are highly recommended to consult Rossi et al. (2005) and Rossi (2014a).

[13] This also holds for the correlations of the parameters and the matrix of price elasticities. The interested reader is referred to the web appendix that allows replicating our results in `R`.

References

Abramson, C., Andrews, R. L., Currim, I. S., & Jones, M. (2000). Parameter bias from unobserved effects in the multinomial logit model of consumer choice. *Journal of Marketing Research*, 37(4), 410–426.

Agresti, A. (2013). *Categorical Data Analysis*, 3rd Ed., New York: John Wiley & Sons.

Ailawadi, K. L., Gedenk, K., Lutzky, C., & Neslin, S. A. (2007). Decomposition of the sales impact of promotion-induced stockpiling. *Journal of Marketing Research*, 44(3), 450–467.

Ailawadi, K. L., Gedenk, K., & Neslin, S. A. (1999). Heterogeneity and purchase event feedback in choice models: An empirical analysis with implications for model building. *International Journal of Research in Marketing*, 16(3), 177–198.

Ailawadi, K. L., & Neslin, S. A. (1998). The effect of promotion on consumption: Buying more and consuming it faster. *Journal of Marketing Research*, 35(3), 390–398.

Ainslie, A., & Rossi, P. E. (1998). Similarities in choice behavior across product categories. *Marketing Science*, 17(2), 91–106.

Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, 35(3), 384–389.

Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32(2), 152–162.

Allenby, G. M., & Ginter, J. L. (1995). Using extremes to design products and segment markets. *Journal of Marketing Research*, 32(4), 392–403.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1), 57–78.

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40(2), 235–243.

Bell, D. R., & Boztug, Y. (2007). The positive and negative effects of inventory on category purchase: An empirical analysis. *Marketing Letters*, 18(1), 1–14.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2), 242–262.

Berry, S. T., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4), 841–890.

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.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1), 79–95.

Braun, M., & Damien, P. (2016). Scalable rejection sampling for Bayesian hierarchical models. *Marketing Science*, 35(3), 427–444.

Bucklin, R. E., & Lattin, J. M. (1991). A two-state model of purchase incidence and brand choice. *Marketing Science*, 10(1), 24–39.

Bronnenberg, B. J., & Vanhonacker, W. R. (1996). Limited choice sets, local price response, and implied measures of price competition. *Journal of Marketing Research*, 33(2), 163–173.

Chintagunta, P., Dubé, J. P., & Goh, K. Y. (2005). Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models. *Management Science*, 51(5), 832–849.

Chintagunta, P., Hanssens, D. M., & Hauser, J. R. (2016). Marketing Science and Big Data. *Marketing Science*, 35(3), 1–2.

Chiou, L., & Walker, J. L. (2007). Masking identification of discrete choice models under simulation methods. *Journal of Econometrics*, 141(2), 683–703.

Currim, I. S., Mintz, O., & Siddarth, S. (2015). Information accessed or information available? The impact on consumer preferences inferred at a durable product E-commerce website. *Journal of Interactive Marketing*, 29(1), 11–25.

Dubé, J. P., Hitsch, G. J., & Jindal, P. (2014). The joint identification of utility and discount functions from stated choice data: An application to durable goods adoption. *Quantitative Marketing and Economics*, 12(4), 331–377.

Dubé, J. P., Hitsch, G. J., & Rossi, P. E. (2010). State dependence and alternative explanations for consumer inertia. *RAND Journal of Economics*, 41(3), 417–445.

Dubé, J. P., Hitsch, G. J., & Rossi, P. E. (2009). Do switching costs make markets less competitive?. *Journal of Marketing Research*, 46(4), 435–445.

Dubé, J. P., Hitsch, G. J., Rossi, P. E., & Vitorino, M. A. (2008). Category pricing with state-dependent utility. *Marketing Science*, 27(3), 417–429.

Elshiewy, O., Zenetti, G., & Boztug, Y. (2017). Differences between classical and Bayesian estimates for mixed logit models: A replication study. *Journal of Applied Econometrics*, 32 (2), 470–476.

Erdem, T., Keane, M. P., & Sun, B. (2008). A dynamic model of brand choice when price and advertising signal product quality. *Marketing Science*, 27(6), 1111–1125.

Fader, P. S., Lattin, J. M., & Little, J. D. (1992). Estimating nonlinear parameters in the multinomial logit model. *Marketing Science*, 11(4), 372–385.

Fiebig, D. G., Keane, M. P., Louviere, J., & Wasi, N. (2010). The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing Science*, 29(3), 393–421.

Frischknecht, B. D., Eckert, C., Geweke, J., & Louviere, J. J. (2014). A simple method for estimating preference parameters for individuals. *International Journal of Research in Marketing*, 31(1), 35–48.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2015). *Bayesian Data Analysis*, 3rd Ed., Chapman & Hall.

Gensch, D. H., & Recker, W. W. (1979). The multinomial, multiattribute logit choice model. *Journal of Marketing Research*, 16(1), 124–132.

Gilbride, T. J., & Lenk, P. J. (2010). Posterior predictive model checking: An application to multivariate normal heterogeneity. *Journal of Marketing Research*, 47(5), 896–909.

Gönül, F., & Srinivasan, K. (1993). Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. *Marketing Science*, 12(3), 213–229.

Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2(3), 203–238.

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

Hauser, J. R., & Koppelman, F. S. (1979). Alternative perceptual mapping techniques: Relative accuracy and usefulness. *Journal of Marketing Research*, 16(4), 495–506.

Hausman, J. A., & Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, 46(2), 403–426.

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

Hensher, D. A., Rose, J., & Greene, W. H. (2005). The implications on willingness to pay of respondents ignoring specific attributes. *Transportation*, 32(3), 203–222.

Hess, S., & Rose, J. M. (2012). Can scale and coefficient heterogeneity be separated in random coefficients models?. *Transportation*, 39(6), 1225–1239.

Hess, S., & Train, K. (2017). Correlation and scale in mixed logit models. *Journal of Choice Modelling*, 23, 1–8.

Hole, A. R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110(3), 203–205.

Horsky, D., Misra, S., & Nelson, P. (2006). Observed and unobserved preference heterogeneity in brand-choice models. *Marketing Science*, 25(4), 322–335.

Huber, J., & Train, K. (2001). On the similarity of classical and Bayesian estimates of individual mean partworths. *Marketing Letters*, 12(3), 259–269.

Jacobs, B. J., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, 35(3), 389–404.

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

Jiang, R., Manchanda, P., & Rossi, P. E. (2009). Bayesian analysis of random coefficient logit models using aggregate data. *Journal of Econometrics*, 149(2), 136–148.

Kalyanam, K. (1996). Pricing decisions under demand uncertainty: A Bayesian mixture model approach. *Marketing Science*, 15(3), 207–221.

Kamakura, W. A., Kim, B. D., & Lee, J. (1996). Modeling preference and structural heterogeneity in consumer choice. *Marketing Science*, 15(2), 152–172.

Kamakura, W. A., & Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26(4), 379–390.

Keane, M. (1997). Current issues in discrete choice modeling. *Marketing Letters*, 8(3), 307–322.

Kim, B. D., Blattberg, R. C., & Rossi, P. E. (1995). Modeling the distribution of price sensitivity and implications for optimal retail pricing. *Journal of Business & Economic Statistics*, 13(3), 291–303.

Krishnamurthi, L., & Raj, S. P. (1988). A model of brand choice and purchase quantity price sensitivities. *Marketing Science*, 7(1), 1–20.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent glass regression in R. *Journal of Statistical Software*, 11(8), 1–18.

Luce, R. (1959). *Individual choice behavior*, New York: John Wiley & Sons.

Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of Marketing Research*, 20(4), 350–367.

Montgomery, D. B., Massy, W. F., & Morrison, D. G. (1970). *Stochastic models of buying behavior*, MIT Press.

McFadden, D. (2001). Economic choices. *American Economic Review*, 91(3), 351–378.

McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (Ed.), *Frontiers in Econometrics*, New York: Academic Press, 105–142.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *RAND Journal of Economics*, 31(3), 395–421.

Park, S., & Gupta, S. (2009). Simulated maximum likelihood estimator for the random coefficient logit model using aggregate data. *Journal of Marketing Research*, 46(4), 531–542.

Petrin, A., & Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1), 3–13.

Revelt, D., & Train, K. (2000). Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier. *Working Paper*. Department of Economics, UCB.

Rossi, P. E. (2014a). *Bayesian non- and semi-parametric methods and applications*, Princeton University Press.

Rossi, P. E. (2014b). Invited paper – Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science*, 33(5), 655–672.

Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). *Bayesian statistics and marketing*. New York: John Wiley & Sons
Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, 22(3), 304–328.
Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4), 321–340.
Rossi, P. E., & Allenby, G. M. (1993). A Bayesian approach to estimating household parameters. *Journal of Marketing Research*, 30(2), 171–182.
Russell, G. (2014). Brand Choice Models. In R. S. Winer, & S. A. Neslin (Eds.), *History of Marketing Science*. World Scientific – Now Publishers Series in Business, 3, 19–46.
Shah, D., Kumar, V., & Zhao, Y. (2015). Diagnosing brand performance: Accounting for the dynamic impact of product availability with aggregate data. *Journal of Marketing Research*, 52(2), 147–165.
Shugan, S. M. (2004). Endogeneity in marketing decision models. *Marketing Science*, 23(1), 1–3.
Silk, A. J., & Urban, G. L. (1978). Pre-test-market evaluation of new packaged goods: A model and measurement methodology. *Journal of Marketing Research*, 15(2), 171–191.
Sudhir, K. (2016). Editorial – The exploration-exploitation trade-off and efficiency in knowledge production. *Marketing Science*, 35(1), 1–9.
Thurstone, L. (1927). Psychophysical analysis, *American Journal of Psychology*, 38(3), 368–389.
Thurstone, L. (1959). *The measurement of values*, Chicago: University of Chicago Press.
Train, K. (2000). Halton sequences for mixed logit. *Working Paper*, Department of Economics, UCB.
Train, K. (2009). *Discrete choice methods with simulation*, 2nd Ed., Cambridge university press.
Train, K. (2016). Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, 19, 40–53.
Train, K., McFadden, D. L., & Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *RAND Journal of Economics*, 18(1), 109–123.
Tversky, A. (1972). Choice by elimination. *Journal of Mathematical Psychology*, 9(4), 341–367.
Van Nierop, E., Bronnenberg, B., Paap, R., Wedel, M., & Franses, P. H. (2010). Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research*, 47(1), 63–74.
Villas-Boas, J. M., & Winer, R. S. (1999). Endogeneity in brand choice models. *Management Science*, 45(10), 1324–1338.
Winer, R. S. (1986). A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research*, 13(2), 250–256.
Yegoryan, N., Guhl, D., & Klapper, D. (2016). Understanding the consequences of attribute non-attendance in discrete choice models. *Working paper*, Humboldt University Berlin.
Zenetti, G., & Otter, T. (2014). Bayesian estimation of the random coefficients logit from aggregate count data. *Quantitative Marketing and Economics*, 12(1), 43–84.

Keywords

Multinomial Logit Models, Brand Choice, Discrete Choice Models, Consumer Heterogeneity.