Goeken, Nils; Kurz, Peter; Steiner, Winfried J.

Multimodal preference heterogeneity in choice-based conjoint analysis: a simulation study

Journal of Business Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Goeken, Nils; Kurz, Peter; Steiner, Winfried J. (2023) : Multimodal preference heterogeneity in choice-based conjoint analysis: a simulation study, Journal of Business Economics, ISSN 1861-8928, Springer, Berlin, Heidelberg, Vol. 94, Iss. 1, pp. 137-185, https://doi.org/10.1007/s11573-023-01156-6

Journal of Business Economics (2024) 94:137–185
https://doi.org/10.1007/s11573-023-01156-6

ORIGINAL PAPER

Multimodal preference heterogeneity in choice-based conjoint analysis: a simulation study

Nils Goeken[1] · **Peter Kurz**[2] · **Winfried J. Steiner**[1]

Accepted: 3 April 2023 / Published online: 26 June 2023

Abstract

The most commonly used variant of conjoint analysis is choice-based conjoint (CBC). Here, hierarchical Bayesian (HB) multinomial logit (MNL) models are widely used for preference estimation at the individual respondent level. A new and very flexible approach to address multimodal and skewed preference heterogeneity in the context of CBC is the Dirichlet Process Mixture (DPM) MNL model. The number and masses of components do not have to be predisposed like in the latent class (LC) MNL model or in the mixture-of-normals (MoN) MNL model. The aim of this Monte Carlo study is to evaluate the performance of Bayesian choice models (basic MNL, HB-MNL, MoN-MNL, LC-MNL and DPM-MNL models) under varying data conditions (especially under multimodal heterogeneity structures) using statistical criteria for parameter recovery, goodness-of-fit and predictive accuracy. The core finding from this Monte Carlo study is that the standard HB-MNL model appears to be highly robust in multimodal preference settings**.**

Keywords Choice-based conjoint analysis · Hierarchical Bayesian estimation · Heterogeneity · Dirichlet Process Mixture · Monte Carlo study

JEL Classification Marketing (M31)

✉ Nils Goeken
nils.goeken@tu-clausthal.de

Peter Kurz
p.kurz@bms-net.de

Winfried J. Steiner
winfried.steiner@tu-clausthal.de

[1] Institute of Management and Economics, Department of Marketing, Clausthal University of Technology, Julius-Albert-Straße 2, 38678 Clausthal-Zellerfeld, Germany

[2] Bms Marketing Research + Strategy, Landsberger Straße 487, 81241 Munich, Germany

1 Introduction

Addressing consumer heterogeneity in choice models is an issue in the marketing literature since the mid-1990s (e.g., Allenby and Ginter 1995; Allenby and Rossi 1998; Rossi et al. 1996). Using appropriate statistical estimation techniques makes it possible for researchers and practitioners to analyze and fully understand markets with truly heterogeneous and/or segment-specific market structures. To date, the most widely applied discrete choice model is the multinomial logit (MNL) model (e.g., Horowitz and Nesheim 2021; Keane et al. 2021), which dates back to McFadden (1973). Considering random taste variation in the MNL model nowadays allows the researcher to derive implications at the individual respondent level and also to avoid or relax the stuck-in-the-middle problem by using individual-level estimates for decisions at the market level (e.g. if a firm plans to launch one new product for an aggregate of consumers). Wedel et al. (1999) distinguished between continuous and discrete representations of consumer heterogeneity. Although the "true" distribution of consumer heterogeneity is often continuous, the concept of the existence of a discrete number of market segments is often more attractive and easier to understand, especially from a managerial point of view (e.g., Ebbes et al. 2015; Tuma and Decker 2013). Whereas discrete approaches often over-simplify the concept of heterogeneity, continuous approaches may not be flexible enough to reproduce consumer heterogeneity adequately, especially if a unimodal heterogeneity distribution is assumed (Allenby and Rossi 1998; Rossi et al. 2005).

Choice models accounting for discrete and continuous representations of heterogeneity became popular for analyzing stated preferences using choice-based conjoint (CBC) data, too (Louviere and Woodworth 1983). On the one hand, the finite mixture MNL approach, proposed by Kamakura and Russell (1989) for the analysis of panel data, was applied to CBC data (DeSarbo et al. 1995; Kamakura et al. 1994; Moore et al. 1998). This approach, also known as latent class (LC) MNL model, divides the market into a manageable number of homogeneous segments with different preference and elasticity structures. On the other hand, Allenby et al. (1995), Allenby and Ginter (1995) and Lenk et al. (1996) published milestone articles for the application of models with continuous representations of heterogeneity to CBC data using hierarchical Bayesian (HB) estimation techniques. Using a normal distribution became the standard procedure to represent preference heterogeneity, referred to as HB-MNL model in the following (e.g., Kim et al. 2007; Webb et al. 2021). A number of researchers have tested and compared the capability and the statistical performance of HB-MNL vs. LC-MNL models, providing ambiguous findings, see e.g. Paetz and Steiner (2017) or Paetz et al. (2019) for detailed reviews. Andrews et al. (2002a) reported that the HB-MNL model worked quite robust even in case of multimodal preference structures. However, it is well known that the thin tails of the normal distribution tend to shrink unit-level estimates toward the center of the data (Rossi et al. 2005). This shrinkage, especially in multimodal data settings, could mask important information (e.g., new or different market structures) (Rossi et al. 2005).

As a generalization of the finite mixture model, the mixture-of-normals (MoN) approach avoids the drawbacks of both the LC-MNL (assumption of homogeneous market segments) and the HB-MNL model (assumption of a unimodal heterogeneity distribution), see Lenk and DeSarbo (2000). Here, a mixture of several multivariate normal distributions representing consumer heterogeneity is applied to a MNL model (Allenby et al. 1998). Using a sufficient number of components, any desired heterogeneity distribution can be approximated using a MoN (e.g., heavy-tailed, multimodal and skewed distributions), see Rossi et al. (2005) or Train (2009). Ebbes et al. (2015) and Chen et al. (2017) more recently reported a better performance of MoN-MNL models in comparison to LC-MNL models in data sets with a large within-segment consumer heterogeneity (Chen et al. 2017) and in the presence of continuous heterogeneity structures (Ebbes et al. 2015).

An additional variant of a discrete choice model for capturing consumer heterogeneity is a hierarchical MNL model with a Dirichlet Process Prior (Voleti et al. 2017). In this way, the researcher is able to model heterogeneity of an unknown form, which allows to classify this approach (as well as the MoN) as Bayesian semi-parametric method (Ansari and Mela 2003; Rossi 2014). One variant of a hierarchical MNL model with a Dirichlet Process Prior is the Dirichlet Process Mixture (DPM) MNL model. Here, part-worth utilities are drawn from continuous distributions (here multivariate normal distributions), where population means and covariances follow a Dirichlet Process. In other words, the continuous distribution is centered around the discrete part-worth utilities of the Dirichlet Process Prior (Voleti et al. 2017). The consideration of within-segment heterogeneity is – as well as in the MoN-MNL – a strength of the DPM-MNL model. Ferguson (1973) and Antoniak (1974) introduced the Dirichlet Process, and although e.g. Escobar and West illustrated Bayesian density estimation based on a Dirichlet Process already in 1995 (Escobar and West 1995), its application in the context of CBC data has been proposed only recently. An advantage of the DPM-MNL is that the number and composition of components are determined as a result a posteriori. Post hoc procedures (e.g., Andrews and Currim 2003) to find the optimal number of segments (components) – like in LC-MNL or MoN-MNL models – are no longer required (Ebbes et al. 2015; Kim et al. 2004; Voleti et al. 2017). Table 1 summarizes the strengths and weaknesses for each of the four model types (LC-MNL, HB-MNL, MoN-MNL, DPM-MNL).

A number of Monte Carlo studies related to conjoint analysis and discrete choice models have been conducted previously, focusing on

- the comparison of different conjoint segmentation methods (Vriens et al. 1996),
- the comparison of different variants of MNL models to capture preference heterogeneity (in particular comparing HB-MNL or MoN-MNL versus LC-MNL models, see Andrews et al. (2002a, 2002b), Otter et al. (2004), Chen et al. (2017), and Ebbes et al. (2015)),
- the comparison of HB-MNL models involving different levels of information (Wirth 2010),
- the analysis of the statistical capabilities of the HB-MNL model for extreme settings of CBC design parameters (Hein et al. 2020), or

Table 1 Strengths and weaknesses of choice models with different representations of preference heterogeneity

Model	Strength	Weakness
LC-MNL	• able to capture segment-specific preference heterogeneity • attractive and easy to understand from a managerial perspective • standard software available	• assumes strictly homogeneous segment preferences • ignores individual preference heterogeneity • number of segments must be pre-specified, making model selection procedures necessary
HB-MNL	• able to capture individual preference heterogeneity • standard software available	• assumption of a single normal heterogeneity distribution • a unimodal distribution is probably not flexible enough to reproduce individual heterogeneity, if clearly separable market segments exist
MoN-MNL	• able to approximate any desired heterogeneity distribution including multimodal, skewed, and/or heavy-tailed distributions • able to simultaneously reproduce market segments and/or individual (within-segment) preference heterogeneity	• model estimation and interpretation of results more complex compared to both HB-MNL and LC-MNL models • number of components must be pre-specified, making model selection procedures necessary (if not the "shut down" procedure of Rossi (2014) is used) • no standard software available
DPM-MNL	• same strengths as MoN-MNL models • optimal number of components (segments) is determined automatically, i.e. no model selection procedure necessary	• model framework still more complex compared to the MoN-MNL model • interpretation of results more complex compared to both HB-MNL and LC-MNL models • no standard software available

- the capability of DPM-MNL models to capture differently shaped heterogeneity distributions (Burda et al. 2008; Li and Ansari 2014).

To the best of our knowledge, no Monte Carlo study has yet systematically explored the comparative performance between LC-MNL, HB-MNL, MoN-MNL and DPM-MNL models for CBC data, with all of these models embedded in the same fully Bayesian estimation framework.

In a study by Voleti et al. (2017), the four models were empirically compared on the basis of eleven CBC data sets. The data sets varied in the number of respondents, the number of choice tasks per respondent, the number of alternatives per task, the number of attributes as well as the number of part-worth utilities to be estimated per respondent (which also depends on the number of attribute levels). The authors focused on the predictive accuracy of the different approaches and found that the DPM-MNL outperformed the competing models in terms of holdout sample hit rates and holdout sample hit probabilities. Importantly, on average, the HB-MNL model provided the second-best predictive performance. More, recently, Goeken et al. (2021) also compared the HB-MNL, the MoN-MNL, and the DPM-MNL models (but not the LC-MNL) in an empirical study, applying them to a real multi-country CBC data set for tires. The authors reported a slightly higher cross-validated hit rate for the DPM-MNL compared to both the MoN-MNL and the HB-MNL, thus confirming the tendency of a better predictive performance of the DPM-MNL in empirical settings. But again, the HB-MNL model was close to the DPM-MNL in its predictive accuracy.

Voleti et al. (2017, p. 334) further stated that the "recovery of parameters is also a relevant objective. However, the only way to address this issue is through computer simulations. […] We leave it to future research to address the issue of parameter recovery under alternative assumptions regarding the true distribution of heterogeneity." Since Goeken et al. (2021) as well focused on empirical data and did not provide any findings for simulated data, we pick up the suggestion of Voleti et al. (2017) in this paper, and study the statistical performance of choice models with different representations of heterogeneity in a Monte Carlo study for CBC data. In particular, we compare the LC-MNL, HB-MNL, MoN-MNL and DPM-MNL models under varying experimental conditions for parameter recovery, goodness-of-fit and predictive accuracy. Like Andrews et al. (2002a), we further incorporate the aggregate MNL model that completely ignores heterogeneity as a benchmark for all heterogeneous models. As opposed to earlier simulation studies, we compare these choice models in *one* Monte Carlo study and estimate all models within the same Bayesian estimation framework.

Parameter recovery is an important criterion for product design decisions as parameters (part-worth utilities in CBC studies) relate to values of product attribute levels and managers are interested to find the best attribute levels for their products. How well a method can recover hidden "true" utility structures can only be studied with artificial data, but knowing which method under which condition is theoretically better in this aspect constitutes an important asset for managers. Independent of whether companies tailor their products to individual customers or not, it is essential and also standard to measure parameter recovery at the individual respondent

level (e.g., Andrews et al. 2002a, 2002b, 2008). In other words, although managers might not be interested in parameter values (preference structures) of specific respondents, a better parameter recovery at the individual respondent level should enable managers to come closer to the real expectations (true preferences) of customers even if product line decisions are subsequently made on a more aggregate level. Market simulations using choice simulators are typically conducted based on individual parameters, even more so as it is well-known that parameter estimates from aggregate models can be strongly biased ("stuck-in-the-middle"). Sometimes, however, companies are also interested in knowing preference parameters of individual respondents, like e.g. in the discrete choice experiments for app-based recommender systems conducted by Danaf et al. (2019). There are also examples for commercial applications where individual-level estimates were the focus, e.g. studies about individual preferences for hair coloration or for preferred products in online shopping trips. Not least, we generally expect personalization efforts of firms and related CBC experiments to further increase in digital environments.

On the other hand, studying the predictive accuracy of the different models under experimental conditions can either generalize the empirical findings of Voleti et al. (2017) or reveal conditions where a different predictive performance can be expected. Predictive accuracy is as well an important dimension for management decisions, since managers are interested in predicting shares of choice (preference shares) as accurate as possible. It has been shown, however, that a model with a high predictive accuracy not necessarily must provide a high accuracy in recovering true parameter values (and the reverse). While minimizing errors in shares-of-choice forecasts represents a natural aggregate measure, it is further also common to assess the predictive validity of a model based on individual-level measures like hit rates or hit probabilities, as used in Voleti et al. (2017). If actual market share data are not available to validate shares of choice predictions, model validation can also be based on individual-level measures (like hit rates in holdout tasks) to find the best model for market simulations. We use the latter approach to provide comparability to Voleti et al. (2017).

To carve out differences in the statistical performance between the classes of models with discrete versus continuous representations of heterogeneity, we specifically vary the levels of within-segment and between-segment heterogeneity. In particular, we want to investigate (1) which representation of heterogeneity is favorable to analyze CBC data, (2) if there is a clear recommendation toward one model for discovering multimodal heterogeneous preference structures and (3) whether (and if how) related findings vary depending on specific levels of our experimental factors. Furthermore, we are particularly interested in (4) how robust the HB-MNL model performs especially in terms of parameter recovery and predictive accuracy compared to the other heterogeneous models due to its underlying unimodal preference distribution which seems least appropriate for segmented markets as considered here. Finally, we want to prove (5) whether the empirical findings of Voleti et al. (2017) with regard to the predictive performance of the models hold for simulated data, too.

In the next section, we propose the design of our Monte Carlo study. In particular, we describe the different choice models, the estimation process, the

performance measures used, and the data generation process including experimental factors and factor levels. We subsequently present the results of the Monte Carlo study, discuss implications and provide an outlook onto future research perspectives. We used the R software (R Core Team 2020) for data generation, choice design construction, model estimation and model evaluation. For model estimation, we used the bayesm package (Rossi 2019) within the R software.

2 Design of the Monte Carlo study

2.1 Models

Since the 1990s, hierarchical Bayesian models have been used for part-worth utility estimation in a CBC framework. The strength of these methods is the ability to yield part-worth utilities at the individual respondent level even when little individual respondent information is available. This is possible by using prior distributions, which borrow information from the sample population (population mean and population covariance). Using a multivariate normal distribution as a first-stage prior has become the state–of–the–art to represent heterogeneity. However, the use of a single normal distribution can be considered as a very conservative approach. Unit-level estimates are shrank toward the population mean, which may mask potential multimodalities in consumer preferences (Rossi et al. 2005). Using a mixture of normal distributions as a first-stage prior can relax this weakness. In particular, multimodal heterogeneity structures as well as thick tails and skewed distributions can be modelled that way. Allenby et al. (1998) pointed out that many distributions can be approximated by using the MoN approach.

Let us denote the utility respondent n ($n = 1, \ldots, N$) obtains from alternative j ($j = 1, \ldots, J$) in choice situation s ($s = 1, \ldots, S$) as

$$U_{njs} = V_{njs} + \varepsilon_{njs}, \tag{1}$$

where $V_{njs} = \beta_n' x_{njs}$ and ε_{njs} represent the deterministic utility and the stochastic utility components, respectively. β_n denotes the vector of part-worth utilities of respondent n, and x_{njs} is a binary coded vector indicating the attribute levels of alternative j offered to respondent n in choice situation s. Assuming that the error term ε_{njs} follows a Gumbel distribution we obtain the MNL model (Train 2009):

$$P_{njs}^{MNL} = \frac{e^{V_{njs}}}{\sum_i e^{V_{nis}}}. \tag{2}$$

To be able to model multimodality with a mixture-of-normals approach consisting of T components, we can specify the hierarchical model as follows (Rossi et al. 2005; Rossi 2014):

$$\begin{aligned}
\beta_n &\sim \mathcal{N}\left(b_{l_n}, W_{l_n}\right), \\
l_n &\sim MN_T(p), \\
p &\sim Dirichlet(\alpha), \\
b_t &\sim \mathcal{N}\left(\bar{b}, w^{-1} W_t\right), \\
W_t &\sim IW(k, \Sigma).
\end{aligned} \tag{3}$$

$l_n \in \{1, \ldots, T\}$ indicates the components from which respondent n can be drawn and follows a multinomial distribution. $p \in \mathbb{R}^T$ denotes the associated probabilities of the multinomial distribution which follow a Dirichlet distribution. $\alpha \in \mathbb{R}^T$ can be interpreted as a tightness parameter, which has an influence on the masses of the components. Rossi (2014) for example shows that larger values of α are associated with a higher prior probability for models with a large number of components. The corresponding population means b_t and the covariance matrices W_t with $t \in \{1, \ldots, T\}$ are normal and inverse Wishart distributed, respectively. The dimensions of b_t and W_t depend on the number of parameters to be estimated. With this model framework, the MoN-MNL model and some nested model variants can be estimated based on CBC data. For $W_{l_n} \neq 0$ and $T = 1$ for example we obtain the HB-MNL model. For diagonal elements of W_{l_n} close to zero we can further approximate[1] the LC-MNL ($T \neq 1$) and the aggregate MNL ($T = 1$) model (Allenby et al. 1998; Lenk and DeSarbo 2000). A reasonable choice of prior settings therefore leads to an approximated LC-MNL and MNL model with a discrete distribution of heterogeneity. By weighting the estimated part-worth utilities of a LC-MNL with the posterior membership probabilities, we obtain part-worth utilities on an individual level (Andrews et al. 2002a).

Using a Dirichlet Process allows for a countable infinite number of components by supplementing the component parameters with additional priors. The DPM-MNL model can therefore be seen as an extension of the MoN-MNL approach. Rossi (2014) comments on a better approximation of multimodal distributions when using Dirichlet Processes. One possible reason for this superiority is that the Dirichlet Process offers the benefits of automatically inferring the number of mixture components. Rossi (2014) stated that in practical applications no more than about 20 components are used in a MoN approach. In some cases this a priori specified number of components in a MoN approach is not near the limiting case (Rossi 2014). Another possible reason for this superiority is that additional priors are placed on the parameters and hyper-parameters of the Dirichlet Process resulting in substantial performance differences and more flexible prior assumptions. To obtain the DPM-MNL model, we replace the Dirichlet prior by a Dirichlet Process:

$$\begin{aligned}
\beta_n &\sim \mathcal{N}\left(b_{l_n}, W_{l_n}\right), \\
\left(b_{l_n}, W_{l_n}\right) &\sim DP\left(\alpha_{DPP}, G_0\right).
\end{aligned} \tag{4}$$

[1] Note that it is not possible to set $W_{l_n} = 0$ (compare Sect. 2.2).

$\alpha_{DPP} \in \mathbb{R}$ is referred to as concentration parameter or Dirichlet Process tightness parameter. Similar to the MoN-MNL, increasing α_{DPP} puts a higher prior probability on models with a large number of components (Rossi 2014). Rossi (2014) chooses a flexible prior[2] for the concentration parameter based on Conley et al. (2008). The advantage of this prior (as compared to e.g. gamma priors) is that the implications for the distribution of the number of possible components are more intuitive to assess (for more details see e.g. Rossi (2014)):

$$p(\alpha_{DPP}) \propto \left(1 - \frac{\alpha_{DPP} - \underline{\alpha}}{\overline{\alpha} - \underline{\alpha}}\right)^{\omega}. \tag{5}$$

Here, $\underline{\alpha} \in \mathbb{R}$ and $\overline{\alpha} \in \mathbb{R}$ are chosen to reflect the range of the probable number of components, and ω is a power parameter. Conley et al. (2008) as well as other authors (e.g., Voleti et al. 2017) describe the modus operandi of the Dirichlet Process and especially of the concentration parameter α_{DPP} with the help of the stick-breaking representation published by Sethuraman (1994). There, the draws from the Dirichlet Process can be represented as an infinite mixture of discrete "atoms" with specific probabilities. Following Conley et al. (2008) the baseline distribution G_0 is parametrized as follows:

$$\begin{aligned} b &\sim \mathcal{N}(0, a^{-1}W), \\ W &\sim IW(\nu, \nu u I). \end{aligned} \tag{6}$$

The priors on a, ν and u are:

$$\begin{aligned} a &\sim U(a_l, a^u), \\ u &\sim U(u_l, u^u), \\ \nu &\sim d - 1 + \exp(z), \\ z &\sim U(\nu_l, \nu^u), \end{aligned} \tag{7}$$

where d is the dimension of the data (here the number of mean part-worth utilities) and U is the uniform distribution. Appropriate prior settings as well as more information on the estimation process are presented in the next section.

2.2 Estimation

In the following, MNL, LC-MNL, HB-MNL, MoN-MNL as well as DPM-MNL models were estimated using Bayesian procedures to obtain part-worth utilities. Markov chain Monte Carlo (MCMC) methods were applied to generate draws from posterior distributions.

[2] Other authors choose a gamma or a uniform prior distribution for the Dirichlet Process tightness parameter. Voleti et al. (2017) stated that the choice of the functional form has only a marginal impact on the number of estimated components.

We used a Gibbs Sampler with a random walk Metropolis step for the MNL coefficients β_n for each respondent n as outlined in Sect. 5.5 of Rossi et al. (2005) and Sect. 5.2 of Rossi (2014). In addition to the "default" prior settings suggested by Rossi (2014),[3] we tested a variety of additional prior settings. We finally adapted the prior settings (in particular the settings for the prior covariance matrix Σ) partly from Sawtooth Software (Sawtooth Software 2016) as they provided the best results in terms of part-worth recovery. Specifically, we chose the following prior configuration to estimate MoN-MNL models (compare Eq. (3)):

$$k = d + 5, w = 0.01, \overline{b} = 0, \alpha = (5, \ldots, 5)^T, \tag{8}$$

where d represents the dimension of the data (here the number of mean part-worth utilities). The prior covariance matrix Σ was chosen according to Sawtooth Software (2016) with a prior variance of 2. Since we can approximate the LC model by the MoN model for diagonal elements of Σ being close to 0 (Allenby et al. 1998), we modified the parameters of the inverse Wishart distribution to estimate the LC models as follows[4]:

$$k = 100, \Sigma = \mathrm{I} \times 0.01, \tag{9}$$

where I is the identity matrix. Note that the prior covariance matrix Σ of the inverse Wishart distribution is a positive-definite matrix. Therefore, we approximate the LC-MNL model by setting the diagonal elements of Σ close to zero. The estimation of both the MoN-MNL and the LC-MNL model was carried out for a fixed number of components $T \in \{1, \ldots, 6\}$, which implicitly included the HB-MNL ($T = 1, W_{l_n} \neq 0$) and the simple MNL model ($T = 1, W_{l_n} = 0$).

To estimate the DPM-MNL model, we set the power parameter ω to 0.8 (Conley et al. 2008). Following Rossi (2014), we set the other prior parameters as follows:

$$a_l = 0.01, a^u = 2, u_l = 0.1, u^u = 4, \nu_l = 0.001, \nu^u = 3. \tag{10}$$

$\underline{\alpha}$ and $\overline{\alpha}$ were chosen to provide a broad prior support for values from 1 to 50 components. We also performed a sensitivity analysis regarding these prior settings and found out that results only differed marginally for different choices of $\underline{\alpha}$ and $\overline{\alpha}$. This is in line with the findings reported by Rossi (2014) and Voleti et al. (2017).

The MCMC sampler was run for 200,000 iterations with a burn-in period of 190,000 iterations. We used only every 50th draw of the remaining 10,000 draws after convergence to reduce autocorrelation among the draws. We evaluated the performance of the various models based on individual draws after the burn-in phase. More precisely, each measure of performance was at first computed on draw-level, and subsequently averaged across the draws. This procedure enables to fully exploit the posterior distribution and also prevents the label switching problem (Frühwirth-Schnatter et al. 2004; Rodríguez and Walker 2014). We monitored the time-series

[3] Rossi (2014) chooses the following prior settings in order to estimate a MoN-MNL model: $k = d + 3, w = 0.01, \overline{b} = 0, \alpha = (5, \ldots, 5)^T, \Sigma = \mathrm{k} \times \mathrm{I}$.

[4] We tested even larger values for k. As a result, the resulting prior became too informative.

plots of parameters and performance measures to ensure convergence of the MCMC chains. We furthermore calculated Gelman and Rubin's potential scale reduction factor to formally prove convergence (Gelman and Rubin 1992). Each check demonstrated that all MCMC chains appeared to reach stable states.

2.3 Experimental design

The choice of the experimental factors and factor levels leans on the Monte Carlo designs used by Vriens et al. (1996), Andrews et al. (2002a), Andrews et al. (2002b) and Andrews and Currim (2003). Overall, six factors were experimentally manipulated in the current study: the model complexity (number of attributes and attribute levels), the number of segments, the separation between segments (between-segment heterogeneity), the segment masses, the degree of within-segment heterogeneity, as well as the number of choice sets to be evaluated per respondent. All factors and their corresponding factor levels used, together with some additional notes, are shown in Table 2, and we will refer back to this table several times in the following.

The more attributes (and attribute levels) are relevant for preference formation, the more parameters (part-worth utilities) a conjoint choice model has. Since attributes in conjoint studies are specified with a discrete number of levels each (including the metric attributes), effects- or dummy-coding is used for parameter estimation (see Sect. 2.1). We vary the number of attributes and levels by analyzing treatments with 6 attributes with 3 levels each, 9 attributes with 4 levels each, or 12 attributes with 5 levels each, leading to choice models with 12, 27, or 48 individual parameters to be estimated for each respondent.[5] A larger number of individual parameters leads to a higher model complexity (factor 1) and given a certain number of choices per respondent to a smaller number of degrees of freedom for model estimation. It can therefore be assumed that a larger number of parameters at the individual respondent level lead to less reliable parameter estimates in all models. Note that we assign no specific meaning to the attributes and do not consider one attribute explicitly to represent the price attribute, since the interpretation of the attributes can be held arbitrary in our Monte Carlo study. Price is often (very) important in empirical studies as well as generally relevant from an economic point of view in CBC studies if related quantities like willingness-to-pay or expected revenue or profit calculations are additionally considered. On the other hand, there are many situations where price is not relevant in choice experiments. Detergents for example have different fragrances and clients are often only interested in preferences for fragrances in combination with the brand (different fragrances do not affect the price of detergents). Smartphone apps are usually not price relevant and can be added along the preferences of the customers. However, apps are relevant for purchase decisions and clients are therefore interested in customers' preferences for apps. Lastly, the development of new cars traditionally goes through a three-stage preference elicitation process: conjoint on design (design clinic), conjoint on features (concept clinic), conjoint on pricing

[5] Note that for L attributes with M levels each, L times (M-1) part-worth utilities are estimated independent whether a dummy- or effects-coding is used.

Table 2 Experimental factors included in the study

Factor	Factor levels	# Factor levels	Remarks
1. Model complexity (# individual parameters)	12, 27, 48	3	6 attributes with 3 levels (12 parameters) 9 attributes with 4 levels (27 parameters) 12 attributes with 5 levels (48 parameters)
2. Number of segments	2, 3, 4	3	Models that allow the reproduction of multiple segments are expected to perform better
3. Separation of segments (between-segment heterogeneity)	small, large	2	The farther apart segment centroids are, the larger the separation of the segments
4. Heterogeneity (within-segment heterogeneity)	small, large	2	Refers to the extent how different consumer preferences within segments are
5. Segment masses	equal, unequal	2	Refers to the number of consumers in each segment (symmetrical or asymmetrical)
6. Number of choice sets per respondent	optimal for estimating main effects, manageable for respondents	2	The optimal number of choice sets allows for an uncorrelated estimation of main effects and depends on the number of attributes and levels (factor 1)

(pricing clinic). In the first stage, price is never included since the "look" of the car is the primary focus here. Price is mostly considered only in the last stage, but some manufactures already additionally conduct a price-only conjoint study in the second stage where preferences for additional features (e.g. color, interior design, entertainment features) are collected.

Given a segmented market structure as assumed in our Monte Carlo study, we expect a better performance of the LC-MNL, MoN-MNL and DPM-MNL models compared to simple MNL and HB-MNL models, since the former models can handle multiple segments (factor 2). The simple MNL is not able to detect any segment structures due to its assumption of parameter homogeneity. Similarly, from a theoretical perspective, the assumption of a unimodal prior in the HB-MNL model is per se not in line with the existence of segmented markets or should make it at least much more difficult to identify existing multimodal preference structures. We therefore expect a worse performance of the HB-MNL model for multimodal preference structures in terms of parameter recovery and prediction accuracy, as well. If segments are less clearly separated from each other (factor 3), i.e. the closer segment centroids are to each other and hence the less between-segment heterogeneity exists, the less distinct the disadvantage for the HB-MNL model is expected to be (e.g., Andrews et al. 2002a).

Including factor 4 allows us to consider more or less (within-segment) heterogeneity in the part-worth utility structures across respondents (Andrews et al. 2002a; Hein et al. 2019; Vriens et al. 1996). Since HB models borrow information from all individuals (respondents) for parameter estimation, the degree of heterogeneity in a sample might affect the individual-level parameter estimates. Previous Monte Carlo studies report different findings about whether models with continuous or discrete representations of heterogeneity are better suited to capture existing preference heterogeneity. While Andrews et al. (2002a, 2002b) have shown that continuous and discrete approaches worked similarly well concerning parameter recovery and predictive validity (Andrews et al. 2002a, 2002b), Otter et al. (2004) reported that the discrete (continuous) approach performed superiorly if the underlying heterogeneity distribution was strictly discrete (continuous). In addition, for sparse data at the individual respondent level, Otter et al. (2004) found the discrete approach to provide a superior parameter recovery and predictive performance.[6]

We expect that both a smaller within-segment and between-segment heterogeneity should positively affect the performance of simple MNL and LC-MNL models, because both only use discrete support points. When the inner-segment heterogeneity is large, we expect a better performance of HB-MNL, MoN-MNL and DPM-MNL models. Furthermore, it can be assumed that it is more difficult to identify the "true" segment structure (factor 2) for more heterogeneous samples, especially if the separation between segments is small (factor 3). The heterogeneity levels are chosen according to Andrews et al. (2002a). Based on a variety of Monte Carlo studies in the context of finite mixture models summarized in a meta-study of Tuma and Decker (2013), we

[6] Note that the Monte Carlo studies of Andrews et al. (2002a), Andrews et al. (2002b) and Otter et al. (2004) also considered both within-segment heterogeneity and between-segment heterogeneity as experimental factors, but did not include the MoN-MNL and DPM-MNL models for comparison.

generated preference structures for 2, 3 and 4 segments. We further expect problems for the LC-MNL model in identifying small segments, i.e. when the masses of components are rather small. In other words, we expect a better performance of LC-MNL models in the symmetric case when the number of respondents is equal across segments compared to the asymmetric case when segment sizes are different from each other (one large, one or several small segments, factor 5) (Andrews and Currim 2003; Dias and Vermunt 2007).

Factor 6 addresses the implementation of CBC studies in market research practice and the related problem that clients want to incorporate more and more attributes while keeping the choice task manageable for respondents (e.g., Hauser and Rao 2004; Hein et al. 2020). In their meta-analyses of empirical CBC studies, both Hoogerbrugge and van der Wagt (2006) and Kurz and Binner (2012) could show that using too many choice tasks per respondent (more than about 15) did no longer increase or may even decrease the predictive performance of HB models, since respondents tend to apply simplification strategies or become disengaged in later choice tasks (also referred to as "individual choice task threshold"). If a choice design comprises more choice tasks than manageable for a single respondent, the researcher can split the design into several versions. Of course, in a Monte Carlo study the number of choice tasks to be completed by a respondent is not relevant as artificial respondents do not become fatigue. Still, by varying the length of the choice task we are able to analyze the statistical effects of shorter-than-optimal designs (regarding the criterion of orthogonality on the individual respondent level) on the model performance. We expect a worse performance of models when splitting the choice task into several versions since then the choice design does not allow an uncorrelated estimation of main effects (Street et al. 2005).

Note that we did not vary factor 6 for treatments with 12 individual parameters (factor 1, see Table 2) since here the resulting optimal number of 18 choice sets per respondent is (nearly) compatible with the "individual choice tasks threshold" of respondents (see next section for more details). Thus, we obtain $2^3 \times 3^2 + 2^4 \times 3 = 120$ experimental data conditions (treatments) and with one replication (i.e., two runs) per treatment 240 data sets. In empirical applications, only the number of parameters to be estimated at the individual respondent level (factor 1) and the number of choice sets respectively the number of versions (factor 6) are observable prior to estimation. In contrast, the number of segments (factor 2), the separation of segments (factor 3), the amount of heterogeneity across respondents (factor 4), and the masses of segments (factor 5) are not known a priori to model estimation. Table 2 provides an overview of all factors and their corresponding factor levels used in our Monte Carlo study together with additional notes.

Generally, we expect the MoN-MNL and especially the DPM-MNL models to outperform the other models (with regard to both parameter recovery, model fit and predictive validity) because they accommodate both within-segment and between-segment heterogeneity. Voleti et al. (2017) only focused on predictive capabilities and found out that the DPM-MNL model can improve the predictive validity. However, for their data sets, the MoN-MNL model was not able to improve the predictive validity over HB-MNL models. On the other hand, data in CBC studies are generally quite sparse on the individual respondent level, and it is therefore not clear whether

the performance of the more complex MoN-MNL and DPM-MNL models necessarily outperforms the more restrictive (single) multivariate HB-MNL model.[7] To the best of our knowledge, no Monte Carlo study related to conjoint data has yet compared the goodness of parameter recovery of MoN-MNL and DPM-MNL models. In particular, we will analyze how well these two types of models are able to detect the "true" part-worth utility structure compared to the other models in extreme scenarios (e.g., 4 segments, small separation, large heterogeneity and asymmetric segment masses). We further controlled for the "overlapping mixtures problem" by holding the sample size constant (Kim et al. 2004), and used 600 respondents following the study of Wirth (2010).

2.4 Data generation

The following section describes how the synthetic data sets were generated in our Monte Carlo study. The data generation process can be divided into the construction of the choice task design, the generation of individual part-worth utilities, and the generation of choice decisions based on the choice task design and individual part-worth utilities. The data sets that support the findings of this study are available from the corresponding author upon request.

2.4.1 Choice task design

Following Street et al. (2005) and Street and Burgess (2007), we constructed optimal choice designs. Determinants for the design generation were the model complexity (factor 1) as well as the number of choice task versions (factor 6). The number of individual parameters, i.e. part-worth utilities to be estimated for each respondent, results from the specification of the number of attributes and attribute levels, as already outlined in the last subsection and summarized in Table 3 below. We used symmetrical designs (i.e., the same number of attribute levels across attributes) to control the number-of-levels effect (Verlegh et al. 2002).

Depending on the number of attributes and attribute levels, we chose an orthogonal array from Kuhfeld (2019) as a starting orthogonal design to fix the first alternative in each choice set. Further alternatives were then added to the first options in each choice set by generating systematic level changes via modulo arithmetic.

As a result, as many pairs of alternatives in a choice set had assigned different levels for each attribute (Street et al. 2005). To ensure an equal distribution of attribute levels (per attribute) across the choice sets (level balance) as well as an equal distribution of attribute levels across alternatives within each choice set (minimal overlap) we constructed choice sets with 3, 4, or 5 alternatives for treatments with 3, 4 or 5 attribute levels (Table 3), respectively (Street and Burgess 2007). The information matrices of our CBC designs were thus diagonal so that estimates of main effects were uncorrelated. By comparing the determinants of the information matrices with the determinants of the information matrices of an optimal design, we

[7] We thank an anonymous reviewer for this note.

Table 3 Model complexity determined by the number of attributes and attribute levels in the CBC design

Individual parameters	Attributes	Levels	Shortcut
12	6	3	A6L3
27	9	4	A9L4
48	12	5	A12L5

obtained a D-efficiency of 100% for each of our generated choice designs.[8] Based on the generated optimal choice tasks each synthetic respondent completed all corresponding choice sets on the one hand. This ensured that all main effects could be estimated completely independently from each other on the individual respondent level. On the other hand, the choice task length of these optimal designs may be far too large for real respondents. We therefore split up the generated optimal choice tasks into several versions (where necessary) in order to limit the number of choice sets per respondent to a manageable number (factor 6). As a consequence, choice designs were no longer optimal on an individual respondent level because desirable properties such as orthogonality or level balance could have been negatively affected by the split. However, since the choice sets were randomly split into several versions, they were at least near-optimal (Street and Burgess 2007). For the treatments with 6 attributes with 3 levels each the starting orthogonal design comprised 18 alternatives, which can be just considered a manageable number. Therefore, a split of the optimal design across respondents was not necessary here. In contrast, treatments including 9 attributes with 4 levels each resulted in an optimal choice design with 32 choice sets. Accordingly, the design was divided into two versions with a length of 16 choice sets each. Similarly, the optimal design for treatments involving 12 attributes with 5 levels each was divided into 5 versions with a length of 20 choice sets each. This procedure ensured desirable properties for optimal or at least near-optimal discrete choice experiments (Street and Burgess 2007). Table 4 summarizes how factor 6 was operationalized depending on the model complexity (factor 1). To be able to assess the predictive validity of the competing models, three additional holdout choice tasks were randomly generated for each respondent.

2.4.2 Part-worth utilities

For each of the 240 data sets (i.e., for each treatment and replication), individual part-worth utilities were generated in such a way that they followed a mixture of multivariate normal distributions. Leaning on Wirth (2010), elements of a vector of initial "true" mean part-worth utilities ($\beta_{start} \in \mathbb{R}^d$, where d is the total number of attribute levels) were drawn from a uniform distribution within the range between -5 and $+5$. Such a range for mean betas is typical for empirical applications (cf. Wirth 2010). We can confirm this finding of Wirth based on an inspection

[8] The previous Monte Carlo studies also used main-effect designs, i.e. no interactions between attributes were considered for the construction of the choice task designs. However, we estimated all choice models with full covariance matrices.

Table 4 Number of choice tasks per respondent (factor 6) depending on the model complexity

Model complexity	Split of the choice design into versions	Number of choice sets per individual
A6L3	no (optimal)	18
A9L4	no (optimal)	32
	yes (manageable)	16
A12L5	no (optimal)	100
	yes (manageable)	20

of a random sample of 250 real-world HB-CBC studies conducted at one of the largest market research institutes worldwide (with 6 to 12 attributes, 3 to 5 attribute levels, and 11 to 15 choice tasks).[9]

The generation of mean part-worth utilities (centroids) for the segments (factor 2) closely follows the studies of Andrews et al. (2002a) and Andrews and Currim (2003) and is based on the generation of a separation vector that controls the distance between the segment centroids. In particular, the separation between segments was manipulated by generating a vector $sep_z \in \mathbb{R}^d$ with $z \in \{1, 2, \ldots, Z\}$ as segment index and $sep_z \sim N(1, 0.1)$ for a small separation and $N(2, 0.2)$ for a large separation (factor 3), see Andrews and Currim (2003). These vectors were then added to the initial vector of "true" mean part-worth utilities to generate the segment-specific centroids (i.e., "true" segment mean part-worth utilities):

$$\beta_z = \beta_{start} + SIGNS_z \times sep_z. \tag{11}$$

$SIGNS_z \in \mathbb{R}^{d \times d}$ denotes a diagonal matrix containing the values -1 and $+1$, each of which were randomly drawn based on a Bernoulli distribution with parameter 0.5. Finally, the generated segment mean part-worth utilities were rescaled to become zero-based, i.e. so that each first level of an attribute constitutes the reference category with a corresponding part-worth utility of zero. Note that multiplying segment-specific part-worth utilities by a constant factor like in Vriens et al. (1996) or Andrews et al. (2002b) also scales the separation of segments. However, Andrews et al. (2002a) demonstrated that such a procedure affects the scale factor of the MNL model, which makes it difficult to assess parameter recovery (Andrews et al. 2002a). Similarly, multiplying the separation vectors sep_z by a constant other than -1 or $+1$ would confound the scale factor of the MNL and thus the sensitivity of respondents, too (Andrews and Currim 2003).

Next, inner-segment heterogeneity (factor 4) was generated by adding quantities to the mean segment part-worth utilities β_z. These quantities were drawn from a multivariate normal distribution with mean vector 0 and covariance matrix $V_{\beta_z} \in \mathbb{R}^{d \times d}$, the latter which was determined by:

$$V_{\beta_z} = v \times \mathrm{I}_{\beta_z}. \tag{12}$$

[9] Vriens et al. (1996) and Andrews et al. (2002b) used a smaller range of -1.7 to +1.7. However, when estimating the DPM-MNL model it turned out that this range was far too small to identify any segments.

$I_{\beta_z} \in \mathbb{R}^{d \times d}$ denotes the identity matrix, and the scalar v controls the degree of inner-segment heterogeneity with either $v = 0.05$ (small heterogeneity) or $v = 0.25$ (large heterogeneity), see Andrews et al. (2002a).[10] In addition, segment masses (factor 5) were defined to be either equal or unequal. In the symmetric case, the relative size of segment z is equal to $1/Z$. In the asymmetric case, the relative size of the largest segment was fixed to $1.5 \times (1/Z)$, while the remaining respondents were split equally across the other segments with relative segment sizes of $(1 - 1.5 \times (1/Z))/(Z - 1)$.

Table 5 shows the resulting segment masses for the symmetric versus asymmetric case depending on the number of segments considered.

2.4.3 Generation of choices

Based on the generated choice task designs and the generated "true" individual part-worth utilities, deterministic utilities $V_{njs} = \beta_n' x_{njs}$ could be at first computed for each respondent for each alternative in each choice set. Stochastic utilities were subsequently computed by adding a Gumbel distributed error term with standard error variance to the deterministic utilities. Simulated choices were obtained by assuming that each respondent chooses the alternative with the highest stochastic utility from a choice set. Based on the simulated choices part-worth utilities were re-estimated by the different models.

2.5 Measures of performance

We estimated 13 different models for each data set: one aggregate MNL model (as benchmark model), one HB-MNL model, one DPM-MNL model, as well as each five LC-MNL and MoN-MNL models with two to six components. Model selection was at first performed for the estimated LC-MNL and MoN-MNL models to determine the appropriate number of segments, respectively. Though "true" preference structures for a maximum of four segments were generated, we decided to estimate LC-MNL and MoN-MNL models for five and six segments in addition to explore the capabilities of the two types of models to find the "true" number of segments. Subsequent to the model selection process where the best LC-MNL and MoN-MNL solutions were retained, we assessed the statistical performance of the five different types of models. This means that a total of 240 (data sets) × 5 (models) = 1,200 observations were subjected to analysis of variance (ANOVA), i.e. the type of model was included as additional factor in the ANOVAs. Following previous Monte Carlo studies (e.g., Andrews et al. 2002a; Hein et al. 2019; Vriens et al. 1996), we evaluated the performance of the competing models in terms of parameter recovery, goodness-of-fit and predictive accuracy. We used three measures for parameter recovery, three measures for goodness-of-fit, and two measures for predictive accuracy. Each performance measure was

[10] We checked for dominant attributes across experimental conditions after having generated the individual part-worth utilities, since one or two attributes with relatively high importance would reduce the potential effects between these conditions. No abnormalities were observed in this regard. We thank an anonymous reviewer for this note.

Table 5 Segment masses for the symmetric versus asymmetric case depending on factor 2

Number of segments	Equal segment masses				Unequal segment masses			
2	300	300			450	150		
3	200	200	200		300	150	150	
4	150	150	150	150	225	125	125	125

computed 200 times based on the 200 individual HB draws that were saved after the burn-in phase (see Sect. 2.2 above) to fully exploit the information of the posterior distribution. Finally, the draw-based scores were averaged to compare the performance of the models along the measures used.

2.5.1 Model selection

For model selection, we computed the marginal likelihood (ML) by means of the Harmonic Mean estimator (Frühwirth-Schnatter 2004; Newton and Raftery 1994; Rossi et al. 2005):

$$\widehat{L}(y|\text{model}) = \left(\frac{1}{R} \sum_{r=1}^{R} \frac{1}{L\left(\widehat{\beta}^r|\text{model}\right)} \right)^{-1}, \tag{13}$$

where $r = 1, \ldots, R$ denotes the r-th draw of the Markov chain used for computing the harmonic mean. The ML penalizes models for complexity, i.e. models with a larger number of parameters get a higher penalty (Frühwirth-Schnatter 2006; Rossi 2014), and it is common practice to prefer more parsimonious models in the model selection process. Following Wirth (2010) and Rossi (2014), we here used the log marginal likelihood (LML) in order to minimize overflow problems. Similar to Elshiewy et al. (2017), we plotted the LML values against the number of components estimated by the LC-MNL or MoN-MNL models and used the "elbow"-criterion for model selection. Furthermore, we examined the more informative sequence plots of the log-likelihood values to identify possible outliers as suggested in Rossi et al. (2005). Note that the approximation of the LML can be influenced by outliers in the vector of log-likelihoods. Following Voleti et al. (2017) and Zhao et al. (2015), we further applied the deviance information criterion (DIC, Spiegelhalter et al. 2002, 2014) and the Watanabe-Akaike information criterion (WAIC, Watanabe 2010) as additional measures for model selection. The latter (WAIC) is closely related to leave-one-out cross-validation, as discussed in Vehtari et al. (2017). Like the LML, DIC and WAIC as well penalize models for complexity. Contrary to these "explicit" model selection procedures (estimating models with a different number of components and selecting the best one), Rossi (2014, p. 29) has suggested to start with a sufficiently large number of components (we here set $T = 6$, see above) and to allow the MCMC sampler to "shut down" a number of the components in the

posterior (also see Goeken et al. 2021 for an application). We also tested this kind of model selection in our Monte Carlo study.

2.5.2 Parameter recovery

Parameter recovery was measured by the Pearson correlation between the generated ("true") and the re-estimated individual part-worths on the individual draw-level. Since Pearson correlations are not interval-scaled, they were rescaled using Fisher's z-transformation prior to computing the mean Pearson correlation across respondents, and retransformed afterwards to their original scale (Hein et al. 2019, 2020).

As a measure of parameter recovery in absolute terms, the root mean square error (RMSE) between "true" (β_{nal}) and re-estimated part-worth utilities $\hat{\beta}^{\mathrm{r}}_{\mathrm{nal}}$ was calculated:

$$\mathrm{RMSE}\left(\hat{\beta}^{r}\right) = \sqrt{\frac{\sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{a}=1}^{\mathrm{A}} \sum_{\mathrm{l}=1}^{\mathrm{L}} \left(\hat{\beta}^{\mathrm{r}}_{\mathrm{nal}} - \beta_{\mathrm{nal}}\right)^{2}}{\mathrm{NAL}}}, \tag{14}$$

where N, A and L refer to the number of respondents, the number of attributes and the number attribute levels.

In addition to the Pearson correlation and the RMSE, we further determined the proportion of "true" part-worth utilities covered by the 95% credible interval of the draws of the posterior distribution, referred to as %TrueBetas (Hein et al. 2020).

2.5.3 Model fit

The percent certainty, the root likelihood, and the in-sample hit rate were used as measures to compare the goodness-of-fit between models. The percent certainty (PC), also referred to as pseudo R^2, McFadden's R^2, or likelihood-ratio-index, compares the likelihood of an estimated (final) model to the likelihood of the null model, i.e. a model without any explanatory variables (Hauser 1978; Ogawa 1987):

$$PC(\hat{\beta}^{r}) = \frac{LL^{r}_{final} - LL_{null}}{-LL_{null}}, \tag{15}$$

where LL^{r}_{final} and LL_{null} denote the log-likelihood of the (final) estimated model based on draw r and the null log-likelihood.

Log-likelihood values were calculated by

$$LL^{r} = \ln\left(L\left(\hat{\beta}^{r}\right)\right) = \sum_{n=1}^{N} \sum_{j=1}^{\mathrm{J}} \sum_{s=1}^{\mathrm{S_n}} Y_{njs} \ln\left(\hat{P}^{r}_{njs}\right), \tag{16}$$

where $\mathrm{S_n}$ denotes the number of choice sets offered to respondent n. Y_{njs} indicates whether respondent n has chosen alternative j from choice set s, and $\hat{P}^{r}_{njs}$ is the choice probability of respondent n for choosing alternative j in choice set s based on draw r. LL_{null} represents the chance likelihood, that means $\hat{\beta}^{r} = (0, \ldots, 0)^{T} \forall r$.

The root likelihood (RLH) is the geometric mean of hit probabilities (e.g., Jervis et al. 2012)

$$RLH(\hat{\beta}^r) = \sqrt[NS_n]{\prod_{n=1}^{N}\prod_{s=1}^{S_n}\prod_{j=1}^{J}\hat{P}^r_{njs}{}^{Y_{njs}}}. \tag{17}$$

A RLH value equal to the reciprocal of the number of alternatives in a choice set (here: $1/J$) corresponds to completely uninformative utilities of all alternatives (i.e., each alternative has the same utility). In other words, the RLH of the null model equals $1/J$.

The in-sample hit rate (IHR) represents the percentage of first choice hits in the estimation sample (e.g., Andrews et al. 2002b; Voleti et al. 2017). The term first choice hit means that the alternative chosen by a respondent from a choice set is assigned the highest deterministic utility based on the re-estimated part-worth utilities. Note that the first choice rule is invariant to the value of the scale parameter of the Gumbel distribution.

2.5.4 Predictive accuracy

The hit rate was further computed for holdout choice sets to assess the predictive accuracy, referred to as holdout sample hit rate (HHR). For this, three holdout choice tasks were randomly generated for each respondent. Further, we computed the root mean square error between the "true" and predicted deterministic utilities (RMSE(V)) for each draw (e.g., Andrews et al. 2002b):

$$\mathrm{RMSE}\left(\hat{V}^{\mathrm{r}}\right) = \sqrt{\frac{\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{J}}\sum_{\mathrm{S}=1}^{\mathrm{S_n}}\left(\hat{V}^{\mathrm{r}}_{\mathrm{njs}} - V_{\mathrm{njs}}\right)^2}{\mathrm{NJS}_n}} \tag{18}$$

V_{njs} and $\hat{V}^{r}_{njs}$ denote the "true" versus predicted deterministic utilities (the latter based on draw-level) for respondent n, alternative j and choice set s, respectively. The number of holdout choice tasks S_n was held constant in all treatments ($S_n = 3$).

3 Results and discussion

3.1 Effects on parameter recovery, fit and predictive accuracy

The impact of the six experimental factors and the type of model (aggregate MNL, HB-MNL, LC-MNL, MoN-MNL and DPM-MNL) on each of the eight measures of performance was examined by analysis of variance for main effects and first-order interaction effects. The ANOVAs were based on a total of 1,200 observations (240

data sets times 5 models) with 1,130 degrees of freedom for error. Prior to that, the best LC-MNL and MoN-MNL solutions were selected in a first attempt by applying the "elbow" criterion to the plots of the LML values versus the number of components (2 to 6). Figure 1 displays examples for selecting the right number of components via the "elbow" criterion. In the refinement subsection (Sect. 3.2), we will discuss the results from applying the DIC, WAIC and the "shut down" procedure suggested by Rossi (2014) for model selection.

Panels A-C show three different scenarios for treatments with 2, 3, or 4 "true" segments where the LC-MNL model was estimated for 2 to 6 segments. In all three scenarios, the "true" number of components was clearly identifiable by means of the elbow criterion. Using the LC-MNL model, we were able to recover the "true" number of segments by the elbow criterion uniquely in 82% of all data sets. Panels D-F show another three plots for treatments with 2, 3, or 4 "true" segments, this time relating to estimations based on the MoN-MNL model (again for 2 to 6 components). The picture is completely different here since in neither case the "true" number of segments is identified. First, no clear elbow is visible each time, rather the LML continues to improve for an increasing number of components. And second, if one dared to recognize an elbow, it would suggest the wrong number of segments in each of the three scenarios.[11] We think plots like the ones in panels D-F are too diffuse to justify a unique solution (i.e., a clear elbow), thus we chose the solution with six components in such cases. In contrast to the LC-MNL model, we were able to identify the "true" number of components via the elbow criterion in only 2% (!) of all data sets when using the MoN approach (in 5 out of 240 data sets). Overall, the model selection process for the MoN models resulted in 67 solutions with five components (28%) and 136 solutions with six components (57%). In another 13% of the cases, wrong solutions with two to four segments were suggested. Note that also the DPM-MNL model returned the "true" number of components in only 14% of all cases (34 out of 240 data sets). The capability of the DPM-MNL model to recover the "true" number of segments was therefore rather disappointing, too.

We further conducted chi-squared tests to assess significant relationships between the experimental factors and the number of re-estimated components (for LC-MNL and MoN-MNL models based on the best solutions determined by model selection by LML). In the cases where we obtained significant results we subsequently analyzed for each respective factor level how often the "true" number of segments could be identified.

It turned out that the "true" number of segments was correctly recovered by the LC-MNL model at all times for treatments with equal segment masses (symmetric case). In contrast, the hit ratio was only 63% for treatments with unequal segment masses (76 out of 120 cases). The number of components suggested by the DPM-MNL depends on the model complexity, the degree of between-segment heterogeneity (separation), and the degree of inner-segment heterogeneity. Fewer components were suggested for treatments with more parameters to be estimated. In particular,

[11] For example, one could think about an elbow for three segments in panel D, however the "true" number of segments was two.

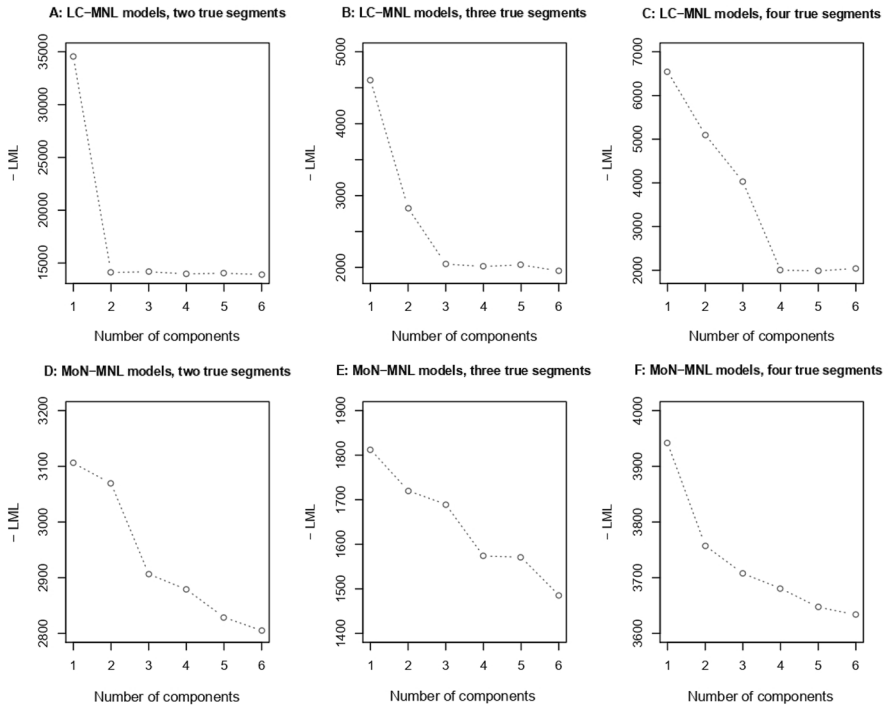

Fig. 1 Selecting the right number of segments via the elbow criterion based on the log marginal likelihood (LML). Panels **A–C** show estimation results for the LC-MNL model, while panels **D–F** refer to estimation results from the MoN-MNL model

a maximum of two segments was found for the treatments with 12 attributes and 5 levels each (81% one-component solutions, 19% two-component solutions). Furthermore, the DPM-MNL models yielded more components in treatments with a small separation between the segments or with a small degree of inner-segment heterogeneity. On the one hand, if the "true" components overlap because they are less clearly separated from each other, the DPM-MNL models tend to a larger number of components. The reason for this result may be that the preference structure of respondents appears more diffuse with less clearly separated segments so that more components are needed to reproduce this diffuse preference pattern. On the other hand, if the "true" components overlap due to a large degree of heterogeneity within segments, the DPM-MNL models tend to a smaller number of components. This may be because preference structures appear to be less multimodal when the "true" segment structures become blurred by a large inner-segment heterogeneity. No significant relationships were found for the MoN-MNL model since 85% of the selected solutions were either 5-component or 6-component solutions. Overall, the LC-MNL model seems to be the best approach by far to recover the "true" number of segments, in particular for scenarios with equal segment masses.

Taking into account only the best LC-MNL and MoN-MNL solutions per data set, we used 1,200 observations (240 data sets times 5 types of models) for analysis of variance.[12] R-squares (adjusted R-squares) range between 0.533 (0.504) and 0.949 (0.946), whereas half of the R squares are larger than 0.9. Most of the main effects (86%) are highly significant ($p < 0.001$), indicating differences in the measures of performance between the corresponding factor levels.

First of all, we recognize that many measures of performance are not significantly affected by the factor segment masses ($p > 0.05$), and if they are (as for the Pearson correlation as well as for IHR and HHR) that F-values turn out rather small compared to other factors. Very high F-values are observed for the type of model which substantially affects all three types of performance measures (recovery, fit, and prediction). In addition, the number of choice sets per respondent represent the factor which most strongly impacts the predictive accuracy (with F-values of 786 and 203 for HHR and RMSE(V)). Higher F-values pointing to substantial differences between factor levels are further observed for the number of parameters in the model (model complexity) and the separation between segments (between-segment heterogeneity).

Furthermore, 62% of the first-order interaction effects are significant. We here, however, consistently observe rather low F-values for nearly all interactions except for some between the type of model and the factor separation (Pearson correlation and the goodness-of-fit statistics). Note that 85% of the interaction effects between the type of model and any of the other factors turn out significant. Here, similar to the main effects, the factor segment masses seems to play a minor role again (as 5 out of 8 interactions between this factor and the type of model are not significant). On the other hand, the separation between segments and the number of parameters in the model (model complexity) are the two factors which most strongly interact with the type of model, in particular w.r.t. goodness-of-fit. It is further noticeable that only 53% of the remaining interaction effects (i.e., excluding interactions where the type of model is involved) are significant. Consequently, the type of model plays a very important role for the goodness of parameter recovery, fit and prediction.

Since even small differences between factor levels may turn out significant for large sample sizes such as in this study ($N = 1200$), we further report related effect sizes measured by Eta square (η^2) in Table 6. Following the guidelines of Cohen (1988), we interpret values of η^2 below 0.06 as small effects, between 0.06 and 0.14 as medium-sized effects, and higher than 0.14 as large effects.

In the following, we concentrate on a more detailed interpretation of factors which show at least medium effect sizes. We observe the largest effect sizes for the type of model with very large effect sizes for all goodness-of-fit measures (> 0.72) and the %TrueBetas measure of parameter recovery, large effect sizes for the Pearson correlation (0.29) and the HHR (0.29), and medium effect sizes for both RMSE measures (0.10). In other words, the effect sizes of the type of model on all performance measures are substantial and most of them are large or very large. We

[12] A summary of F-Tests for main and interaction effects including p-values and R-squares for each performance measure can be found in Table 9 in the Appendix.

further observe medium effect sizes (a) for the number of parameters in the model (model complexity) on parameter recovery (0.10 for both Pearson correlation and RMSE) and the HHR (0.11), (b) for the separation between segments on the Pearson correlation (which measures parameter recovery in relative terms), and (c) for the number of choice sets per respondent on both predictive accuracy statistics, the HHR (0.11) and RMSE(V) (0.08). That also means that 75% of the effect sizes for main effects can be classified as small, and more than half of them are below 0.01. All interaction effects where the type of model is not involved show small if not (as in the very most cases) negligible effect sizes near zero. Considering interactions where the type of model is involved only few (8 out of 48, 17%) show medium-sized effects, and these with only one exception relate to interactions of the type of model with the separation between segments or the model complexity. In particular, we observe medium-sized interactions between (a) the type of model and the separation between segments on the Pearson correlation, PC, RLH, and the HHR, between (b) the type of model and the number of parameters in the model on the RMSE (which measures parameter recovery in absolute terms) and both predictive validity measures (HHR, RMSE(V)), and between (c) the type of model and the degree of inner-segment heterogeneity on RMSE(V). Altogether, the effect sizes provide a rather clear picture: the type of model in particular, and further the number of parameters in the model (model complexity), the separation between segments, as well as the number of choice sets per respondent seem to be the primary drivers for the model performance, while the number of segments, the inner-segment heterogeneity (except for the one interaction effect), and the segment masses do not show any noticeable and in most cases even negligible effect sizes on the model performance. Obviously, the model performance is not substantially affected by the number of segments, although the aggregate MNL and the HB-MNL are not at all or only conditionally able to recover segments. It was further not expected that the degree of inner-segment heterogeneity plays such a weak role especially for the goodness of-parameter recovery.[13]

Table 7 provides the means of the eight performance measures for each individual factor level and further reports significant differences between factor levels based on post hoc tests. For the post hoc tests, we applied the Bonferroni correction to control for the family-wise error rate. For interpreting factor level differences, we again focus on factors which show at least a medium effect size. Rather surprisingly, the HB-MNL model performs excellently in terms of parameter recovery. While, except for the aggregate MNL model, Pearson correlations are comparable across models with high values above 0.95, the HB-MNL model shows a much better performance with regard to the RMSE measure. Especially the DPM-MNL model and the MoN-MNL model perform considerably worse here, showing much larger absolute deviations between the "true" and re-estimated part-worth utilities (DPM-MNL: 3.091; MoN-MNL: 2.608) than the LC-MNL (2.159), the aggregate MNL (2.466) and the HB-MNL (1.650) models. Concerning the percentage of "true" part-worth utilities that lie in the

[13] Andrews et al. (2002b) also observed negligible effects of the factors segment masses and inner-segment heterogeneity on the model performance. However, they did not consider MoN-MNL and DPM-MNL models in their model comparisons.

Table 6 Effect sizes[a] for main and interaction effects on parameter recovery, goodness-of-fit and predictive accuracy measured by Eta squared (η^2). Note that performance measures were calculated based on 200 individual draws and that only the best LC-MNL and MoN-MNL solutions as provided by the model selection (based on the log marginal likelihood) were included in the ANOVAs

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Model (4)	***0.287***	**0.095**	***0.861***	***0.726***	***0.751***	***0.749***	***0.289***	**0.098**
Model complexity (2)	**0.096**	**0.098**	0.017	0.003	0.038	0.030	**0.109**	0.024
Number of segments (2)	0.012	0.001	0.002	0.011	0.008	0.010	0.019	0.004
Separation (1)	**0.067**	0.010	0.000	0.018	0.009	0.006	0.040	0.020
Heterogeneity (1)	0.001	0.022	0.011	0.003	0.006	0.008	0.004	0.014
Segment masses (1)	0.004	0.000	0.000	0.000	0.000	0.001	0.002	0.001
Number of choice sets per respondent (1)	0.057	0.053	0.002	0.002	0.002	0.001	**0.114**	**0.080**
Model × Model complexity	0.041	**0.104**	0.019	0.022	0.032	0.021	**0.077**	**0.108**
Model × Number of segments	0.026	0.012	0.003	0.023	0.013	0.020	0.021	0.014
Model × Separation	**0.128**	0.052	0.005	**0.102**	**0.068**	0.058	**0.110**	0.053
Model × Heterogeneity	0.001	0.047	0.018	0.003	0.002	0.002	0.001	**0.063**
Model × Segment masses	0.005	0.000	0.000	0.000	0.000	0.003	0.002	0.000
Model × Number of choice sets per respondent	0.009	0.002	0.003	0.003	0.007	0.008	0.012	0.006
Model complexity × Number of segments	0.003	0.004	0.001	0.001	0.001	0.001	0.009	0.005
Model complexity × Separation	0.004	0.002	0.000	0.002	0.001	0.001	0.004	0.003
Model complexity × Heterogeneity	0.022	0.016	0.001	0.007	0.007	0.008	0.003	0.033
Model complexity × Segment masses	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.000
Model complexity × Number of choice sets per respondent	0.001	0.001	0.000	0.000	0.000	0.000	0.004	0.008
Number of segments × Separation	0.007	0.000	0.000	0.001	0.000	0.001	0.002	0.003
Number of segments × Heterogeneity	0.001	0.006	0.001	0.000	0.000	0.000	0.001	0.005
Number of segments × Segment masses	0.002	0.001	0.000	0.000	0.000	0.001	0.001	0.000
Number of segments × Number of choice sets per respondent	0.001	0.001	0.000	0.000	0.000	0.000	0.001	0.004

Table 6 (continued)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Separation × Heterogeneity	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.001
Separation × Segment masses	0.002	0.001	0.000	0.000	0.000	0.000	0.001	0.000
Separation × Number of choice sets per respondent	0.006	0.001	0.000	0.000	0.000	0.000	0.005	0.004
Heterogeneity × Segment masses	0.004	0.000	0.000	0.000	0.000	0.000	0.002	0.000
Heterogeneity × Number of choice sets per respondent	0.005	0.000	0.000	0.002	0.002	0.003	0.002	0.000
Segment masses × Number of choice sets per respondent	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000

a: Bold and bold/italicized values indicate effects of medium and large size, respectively

Table 7 Means of performance measures by experimental condition (i.e., at the individual factor level)[a]. Note that performance measures were calculated based on 200 individual draws and that only the best LC-MNL and MoN-MNL solutions as provided by the model selection (based on the log marginal likelihood) were included in the ANOVAs

Factor	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Model								
(1) DPM-MNL	$0.964^{4,5*}$	$3.091^{2,3,4,5*}$	$0.623^{2,3,4,5*}$	$\underline{0.908}^{2,3,4*}$	$\underline{0.879}^{2,3,4*}$	$\underline{0.947}^{2,3,4*}$	$0.824^{3,4,5*}$	$7.469^{2,3,4,5*}$
(2) HB-MNL	$\underline{0.968}^{4*}$	$\underline{1.650}^{1,3,4,5*}$	$\underline{0.746}^{1,3,4,5*}$	$0.888^{1,3,4*}$	$0.855^{1,3,4,5*}$	$0.936^{1,3,4*}$	$\underline{0.832}^{4*}$	$\underline{4.199}^{1,3,4,5*}$
(3) LC-MNL	$\underline{0.967}^{4*}$	$2.159^{1,2,5*}$	$0.014^{1,2,5*}$	$0.828^{1,2,4,5*}$	$0.785^{1,2,4,5*}$	$0.905^{1,2,4,5*}$	$\underline{0.835}^{1,4*}$	$5.395^{1,2,4,5*}$
(4) MNL	$0.926^{1,2,3,5*}$	$2.466^{1,2*}$	$0.021^{1,2,5*}$	$0.583^{1,2,3,5*}$	$0.567^{1,2,3,5*}$	$0.787^{1,2,3,5*}$	$0.765^{1,2,3,5*}$	$6.275^{1,2,3*}$
(5) MoN-MNL	$\underline{0.971}^{1,4*}$	$2.608^{1,2,3*}$	$0.585^{1,2,3,4*}$	$\underline{0.903}^{3,4*}$	$\underline{0.873}^{2,3,4*}$	$\underline{0.943}^{3,4*}$	$\underline{0.838}^{1,4*}$	$6.352^{1,2,3*}$
Model complexity								
(1) A6L3	$0.962^{2,3*}$	2.869^{3*}	0.386^{3*}	0.825	$\underline{0.832}^{2,3*}$	$\underline{0.924}^{2,3*}$	$\underline{0.837}^{2*}$	$\underline{5.426}^{2*}$
(2) A9L4	$0.948^{1,3*}$	2.752^{3*}	0.352^{3*}	0.830	$0.801^{1,3*}$	$0.906^{1,3*}$	$0.798^{1,3*}$	$6.591^{1,3*}$
(3) A12L5	$\underline{0.970}^{1,2*}$	$\underline{1.801}^{1,2*}$	$\underline{0.450}^{1,2*}$	0.813	$0.763^{1,2*}$	$0.891^{1,2*}$	$\underline{0.830}^{2*}$	$\underline{5.540}^{2*}$
Number of segments								
(1) 2	$\underline{0.964}^{3*}$	2.441	0.375	$\underline{0.843}^{2,3*}$	$\underline{0.809}^{3*}$	$\underline{0.913}^{2,3*}$	$\underline{0.827}^{2,3*}$	5.848
(2) 3	<u>0.959</u>	2.313	0.403	0.816^{1*}	<u>0.786</u>	0.900^{1*}	0.818^{1*}	5.729
(3) 4	0.955^{1*}	2.431	0.415	0.807^{1*}	0.780^{1*}	0.897^{1*}	0.810^{1*}	6.236
Separation								
(1) Small	$\underline{0.967}^{*}$	$\underline{2.241}^{*}$	0.394	$\underline{0.841}^{*}$	$\underline{0.805}^{*}$	$\underline{0.909}^{*}$	$\underline{0.829}^{*}$	$\underline{5.443}^{*}$
(2) Large	0.951^{*}	2.549^{*}	0.401	0.803^{*}	0.779^{*}	0.898^{*}	0.808^{*}	6.433^{*}
Heterogeneity								
(1) Small	0.959	2.625^{*}	0.363^{*}	0.830	$\underline{0.802}^{*}$	$\underline{0.910}^{*}$	$\underline{0.822}^{*}$	6.346^{*}
(2) Large	0.960	$\underline{2.165}^{*}$	$\underline{0.433}^{*}$	0.814	0.781	0.898^{*}	0.815^{*}	$\underline{5.530}^{*}$

Table 7 (continued)

Factor	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Segment masses								
(1) Unequal	0.961*	2.373	0.395	0.823	0.793	0.906	0.821	5.821
(2) Equal	0.957*	2.417	0.400	0.821	0.790	0.901	0.816	6.055
Number of choice sets per respondent								
(1) Optimal	0.965*	2.208*	0.406	0.818	0.794	0.905	0.834*	5.114*
(2) Manageable	0.950*	2.675*	0.385	0.828	0.788	0.901	0.795*	7.173*

a: Superscripts on means refer to the factor levels and indicate significant differences at the 0.05 level (according to Bonferroni correction). The underlined values identify the superior conditions.

corresponding 95% credible intervals of the draws, we observe that models with a discrete representation of heterogeneity perform inferior and provide inacceptable results (LC-MNL: 0.014, aggregate MNL 0.021). But again, the HB-MNL model (0.746) performs markedly better than DPM-MNL (0.623) and MoN-MNL (0.585) models.

We observe similar results when comparing the predictive accuracy between the models. Except for the MNL model, the HHR between models differ only marginally with values around 83%. The absolute deviations between "true" and re-estimated total utilities of alternatives are much larger for the DPM-MNL (7.469), the MoN-MNL (6.352) and the aggregate MNL model (6.275) than for the HB-MNL (4.199) and the LC-MNL models (5.395). Again, the HB-MNL model here provides the lowest errors. The better performance of the HB-MNL and LC-MNL models in predicting total utilities of alternatives (RMSE(V)) corresponds with the better performance of both models in terms of absolute errors with regard to parameter recovery (RMSE).

We further observe that DPM-MNL and MoN-MNL models provide the best model fit with respect to all three fit statistics (PC, RLH, IHR), whereas the aggregate MNL model performs by far worst. That the aggregate MNL model performs so much worse here compared to the other four models seems to be the reason for the very large effect sizes of the type of model on the three model fit measures (this also applies to the %TrueBetas measure and in alleviated form to Pearson correlations and holdout sample hit rates, where the aggregate MNL is inferior while the other models perform comparable). On the other hand, including the aggregate MNL model in the ANOVAs provided evidence that it performs not worse than the MoN-MNL model or even significantly better than the DPM-MNL model regarding absolute errors in both parameter recovery and prediction accuracy, respectively. We later check if or how much the ANOVA results for the type of model change when the aggregate MNL model is removed from the analyses (see refinements, Sect. 3.2).

Moreover, optimal choice designs on an individual respondent level enable significantly better predictions compared to the case where respondents evaluate only a smaller (manageable) number of choice sets. We further observe slightly higher Pearson correlations for a smaller separation between segments. In addition, we recognize the best parameter recovery both in relative (Pearson correlation) and absolute (RMSE) terms (and also for the %TrueBetas measure) for the most complex treatment with 12 attributes with 5 levels each (A12L5). Similarly, we also observe a very high HHR (0.83) for the most complex treatment (A12L5), which is markedly higher than for the less complex treatment with 9 attributes with 4 levels (A9L4). We will discuss this in more detail next.

We did not expect these results but rather that a higher number of parameters in the model should lead to a worse parameter recovery and a worse prediction accuracy. However, remember that in our design setup a larger number of attribute levels (3, 4, or 5) not only increased the model complexity but also involved larger choice sets containing more alternatives (e.g., 5 alternatives in treatments with 5 attribute levels, while only 3 alternatives in treatments with 3 attribute levels) as well as a much higher optimal number of choice sets to be evaluated by a respondent (compare Table 4). On the one hand, a higher number of alternatives per choice set makes it more difficult to predict respondents' "true" choices correctly, which should decrease the HHR. On the other hand, more choice sets for each respondent lead to more information on an

individual level, which in turn should improve parameter recovery and prediction accuracy. Obviously, the much larger optimal number of choice sets (100 per respondent) for the most complex treatment (A12L5) compared to the two other treatments (A6L3: 18 choice sets per respondent; A9L4: 32 choice sets per respondent) favors the good performance with regard to parameter recovery and prediction accuracy.

To fully understand the effects of (a) the number of parameters in the model, (b) the level of inner-segment heterogeneity and (c) the separation between segments on the performance measures, it is helpful to examine their interaction effects with the type of model. As before, we only focus on interaction effects which showed an at least medium effect size (see Figs. 2 and 3).

Considering the interaction effects between model complexity and type of model (Fig. 2, panels A-C), we observe that the aggregate MNL model has by far the lowest HHR (panel A). This in particular applies to the most complex treatment (A12L5) where the HHR is about 10% lower than for all competing models (panel A). The corresponding interaction effects on absolute prediction errors (RMSE(V)) and absolute parameter recovery errors (RMSE) show similar patterns (panels B and C). For the treatments with 6 attributes with 3 levels (A6L3) and 9 attributes with 4 levels (A9L4) the aggregate MNL, the LC-MNL and the HB-MNL models perform almost equally well (with slight advantages for the HB-MNL model). For the most complex treatment with 12 attributes with 5 levels (A12L5) the HB-MNL clearly outperforms all other models, whereas the aggregate MNL and the LC-MNL models perform worst here. In addition, we observe that for the less complex treatments (A6L3, A9L4) both absolute error measures turn out very large for the MoN-MNL and DPM-MNL models. The number of re-estimated components (for the MoN-MNL and the

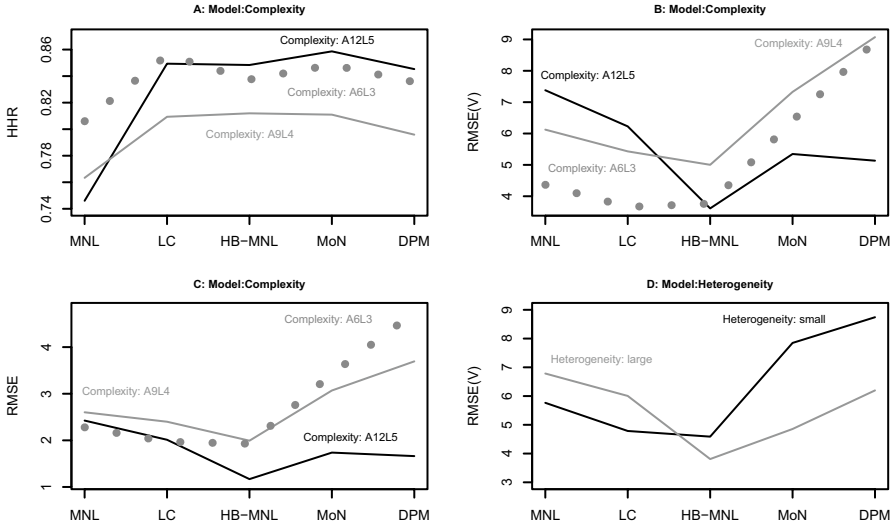

Fig. 2 Panels **A**–**C**: Interaction effects between model complexity and type of model on parameter recovery (RMSE) and prediction accuracy (holdout sample hit rate, RMSE(V)). Panel **D**: Interaction effect between inner-segment heterogeneity and type of model on prediction accuracy (RMSE(V))

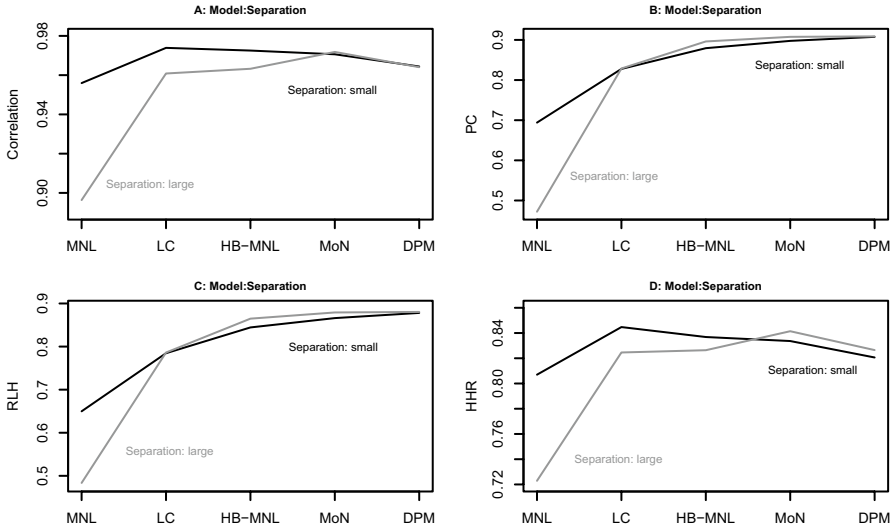

Fig. 3 Interaction effects between separation of segments and type of model on parameter recovery (Pearson correlation), model fit (PC, RLH), and prediction accuracy (holdout sample hit rate)

DPM-MNL models) seems to play a negligible role for the most complex treatment (A12L5). For the DPM-MNL model a maximum of two segments was found for the treatments with 12 attributes and 5 levels each. The MoN model overestimates the number of "true" segments in most cases (see above). However, both models show similar absolute errors for the most complex treatment (lower than the absolute errors for the less complex treatments but higher compared to the HB-MNL model).

A very similar pattern is found for the interaction effect between inner-segment heterogeneity and type of model on absolute prediction errors (RMSE(V)), see panel D in Fig. 2. Here, for treatments with a low inner-segment heterogeneity, the aggregate MNL, the LC-MNL and the HB-MNL models perform again almost equally well (once more with slight advantages in favor of the HB-MNL model), while the MoN-MNL and DPM-MNL models provide inacceptable large prediction errors. As discussed above, this result may be associated with the finding that DPM-MNL (and also MoN-MNL) models tend to more components for treatments with a smaller inner-segment heterogeneity. For treatments with a high inner-segment heterogeneity, the HB-MNL once more clearly outperforms all other models.

When examining the interactions between the separation of segments and the type of model (Fig. 3) on Pearson correlations (parameter recovery), PC, RLH (model fit) and HHR, the most noticeable point is that the aggregate MNL model doesn't work competitively, in particular not if segments are clearly separated from each other. For the treatments with a large separation, all models with continuous representations of heterogeneity (HB-MNL, MoN-MNL, DPM-MNL) perform nearly equally well. The LC-MNL model performs slightly worse in terms of goodness-of-fit (PC, RLH) but is competitive in terms of Pearson correlations and HHRs. Rather similar results can be observed for treatments with a small separation between segments

Table 8 Overview of main results

Factor	Effect size	General tendencies
Model	• Large: parameter recovery (Correlation, %TrueBetas) • Medium: parameter recovery (RMSE) • Large: all goodness-of-fit measures • Large: predictive accuracy (HHR) • Medium: predictive accuracy (RMSE(V))	The HB-MNL model • performs excellently in terms of parameter recovery (RMSE, %TrueBetas) and predictive accuracy (RMSE(V)) • performs similarly compared to the LC-MNL and the MoN-MNL model in terms of correlation (parameter recovery) and HHR (predictive accuracy) DPM-MNL and MoN-MNL models • are superior in terms of goodness-of-fit
Model complexity	• Medium: parameter recovery (Correlation, RMSE) • Medium: predictive accuracy (HHR)	Increasing the model complexity • improves parameter recovery (Correlation, RMSE) • decreases predictive accuracy (HHR)
Number of segments	• Small or near zero: for nearly all performance measures	Due to small effect sizes the number of segments does not seem to substantially affect the model performance
Separation	• Medium: parameter recovery (Correlation)	Increasing the separation • decreases parameter recovery (Correlation)
Heterogeneity	• Small or near zero: for nearly all performance measures	Due to small effect sizes the degree of inner-segment heterogeneity does not seem to substantially affect the model performance
Segment masses	• Small or near zero: for nearly all performance measures	Due to small effect sizes the segment masses do not seem to substantially affect the model performance
Number of choice sets per respondent	• Medium: all predictive accuracy measures	An optimal number of choice sets per respondent • improves predictive accuracy

with the exception that the aggregate MNL doesn't perform such bad here or even comparable to the other models in terms of Pearson correlations and HHRs. For the same two performance measures, the LC-MNL model shows even a slightly better performance than the models with continuous representations of heterogeneity.

We summarize the main results about effect sizes and the impacts of factor levels on the model performance in Table 8.

3.2 Refinements

We performed sensitivity analyses to check if our ANOVA results stayed robust for two differing scenarios.[14] First, we excluded the aggregate MNL model as benchmark model from all ANOVAs. We did this check due to the huge effect sizes (>0.7) we observed for effects of the type of model on all goodness-of-fit statistics (PC, RLH, IHR) and the %TrueBetas measure of parameter recovery (cf. Table 6). The ANOVA results changed only little, providing strong evidence that our findings for the different models are highly robust. The huge effects sizes for the type of model on all goodness-of-fit measures are lower than those reported in Table 6 but still large (>0.44). The effect sizes for the type of model on the Pearson correlation and HHR turn out only small after removing the MNL model from the ANOVAs. Further, we now observe medium effect sizes for the model complexity on RLH and on IHR, medium effect sizes of the degree of within-segment heterogeneity on all goodness-of-fit measures, and medium to large effect sizes for the number of choice sets per respondent on the RMSE (0.06) and on the Pearson correlation (0.185). For the latter factor, corresponding correlations are still high (optimal number of choice sets: 0.975; manageable number of choice sets: 0.957). As expected (cf. Figure 3) the interaction effects between the type of model and the separation between segments on the Pearson correlation, PC, RLH, and HHR become negligible. As mentioned above, it was not expected that the degree of inner-segment heterogeneity plays such a weak role, especially for the goodness of-fit measures. After dropping the MNL model, we now observe medium effect sizes of the factor heterogeneity on goodness of-fit measures (but only on goodness-of-fit measures). A small heterogeneity enables a significantly better fit compared to a large heterogeneity.

Second, we re-estimated all LC-MNL and MoN-MNL models for the given "true" number of segments instead of determining the best solutions by model selection (e.g., see Vriens et al. 1996). As a result, the MoN-MNL model now comes up with much lower absolute errors of parameter recovery (RMSE: 1.826) and prediction accuracy (RMSE(V): 4.518) as well as with a strongly improved %TrueBetas measure of parameter recovery. All other results regarding effect sizes of the experimental factors and the means of performance measures by experimental condition remained extremely robust. Under this approach, the HB-MNL model and the MoN-MNL model work similarly effective in terms of parameter recovery, goodness-of-fit and predictive accuracy. In particular, the MoN-MNL model reveals slight advantages over the HB-MNL

[14] The complete results of the sensitivity analyses are available on request from the authors.

model with respect to Pearson correlations (HB-MNL: 0.968, MoN-MNL: 0.978) and the HHR (HB-MNL: 0.832, MoN-MNL: 0.851), whereas the HB-MNL model performs better w.r.t. the %TrueBetas measure of parameter recovery (HB-MNL: 0.746, MoN-MNL: 0.689). At this point, it is important to note that the "true" number of segments or components is not known in empirical studies.

As noted in Sect. 2.5, we also applied the DIC, WAIC and the "shut down" procedure suggested by Rossi (2014) as alternatives for model selection in addition to the LML criterion. For the LC-MNL model, the true number of segments could be identified as well in the majority of cases when using the DIC (87%) or the WAIC (78%), compared to 82% before via the LML. For the MoN-MNL model, the recovery rate could be improved by using the DIC (9%) or the WAIC (26%), compared to only 2% before (LML). Still, the ability of the MoN-MNL model to identify true segment structures remains quite modest. A somewhat different picture results from using the "shut down" procedure of Rossi (2014) for model selection. For the LC-MNL model, the "true" number of segments could be recovered for only 15% of the simulated data sets, while at least in 38% of all cases by the MoN-MNL model. One possible reason for the rather poor performance of the "shut down" variant obtained for the LC-MNL model could be that the prior configurations used in this paper are a bit more informative than those suggested by Rossi (2014), still they are uninformative. Interestingly, the eight performance measures are hardly affected by the choice of the model selection procedure and remained highly stable for the different model types, as displayed in Table 10 in the Appendix.[15]

Allenby and Rossi (1998) have already noted that the posterior means of individual-level parameters do not have to follow a normal distribution even if the heterogeneity model is represented by the single normal distribution, as in the HB-MNL model. The rationale behind this is that the single normal distribution is only part of the prior, and the posterior is affected by the individual respondent data (cf. Allenby and Rossi 1998, p. 71). Thus, the distribution of the individual-level parameters could be multimodal even if the heterogeneity model is wrong, which would explain – at least to some degree – the very good performance of the HB-MNL model in our study.[16] The following density plots displayed in Fig. 4 should bring more light into this issue, showing for selected treatments how multimodal the simulated preference distributions were and how well individual-level parameters were recovered. In addition, Fig. 4 provides examples for treatments when MoN-MNL and DPM-MNL models overestimated the "true" number of components.

Shown are selected individual-level preference distributions for true versus re-estimated part-worth utilities for different numbers of segments and different combinations of factor levels regarding the factors separation of segments (between-segment heterogeneity) and within-segment heterogeneity. The solid black lines refer to

[15] Note that the eight performance measures are highly robust against the type of model selection with regard to all other experimental factors, too. The corresponding results are available on request from the authors. Also note that different from Voleti et al. (2017) we used the DIC instead of the DIC3 criterion since we have no missing data.

[16] We thank an anonymous reviewer for this note.

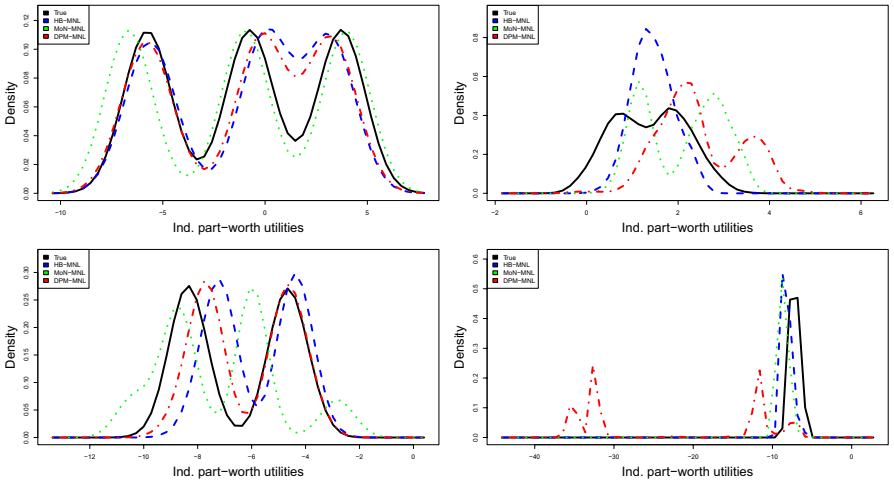

Fig. 4 Selected density plots for "true" distributions of part-worth utilities (black lines) versus re-estimated distributions of part-worth utilities by model type (HB-MNL: blue lines; MoN-MNL: green lines; DPM-MNL: red lines). Upper left panel: 3 "true" segments, large separation, small heterogeneity. Upper right panel: 2 segments, small separation, large heterogeneity. Lower left panel: 2 segments, large separation, large heterogeneity. Lower right panel: 2 segments, small separation, small heterogeneity (note that the two true segments are not visible in the lower right panel due to their small separation and the coarse scaling required to represent the estimated components from the DPM-MNL model; for a finer resolution, see the bottom part of Fig. 5 in the Appendix where the DPM-MNL has been excluded)

the generated "true" preference distributions, while the dashed and/or dotted lines refer to the re-estimated part-worth distributions obtained from the HB-MNL (blue), MoN-MNL (green), and DPM-MNL (red) models.

The upper left panel shows a treatment with three "true" segments, a large separation between segments, and a small individual heterogeneity within these segments. Here, we observed that the MoN-MNL worked well in capturing the three segment structure (one of the few examples where the MoN-MNL performed fine), but that especially the HB-MNL did as well a very good job. Obviously, the posterior means are not constrained to follow the upper level single normal distribution in the HB-MNL model, and can reproduce the true 3-segment structure by adapting to the individual multimodal preference data under this factor level condition. A similar result was found for the treatment with two "true" segments, a *small* separation between segments, and large within-segment heterogeneity, see the upper right panel. Under this condition, the good performance of the HB-MNL model seems more plausible, since the two segments strongly overlap due to the small separation and the large within-segment heterogeneity.

The lower left panel shows a treatment with again two "true" segments, but both a large separation between segments and large within-segment heterogeneity. In this situation, the MoN-MNL suggests a 4-segment solution and therefore overestimates the true number of segments by two components (based on the LML criterion used for model selection here). In contrast, both the HB-MNL and DPM-MNL models very closely recovered the two segments. Finally, the density plots in the lower right panel refer to a treatment with two "true" segments, and both a *small* separation between

segments and *small* within-segment heterogeneity. Here, the DPM-MNL yielded a solution with six components (only four are clearly visible), indicating a clear overfitting. Note that the DPM-MNL models tended to more components in treatments with a small separation between segments or a small extent of inner-segment heterogeneity, compare Sect. 3.1. For a detailed consideration of interaction effects between the factors separation of segments and within-segment heterogeneity, see Figs. 2 and 3.[17]

4 Conclusions, managerial implications, and outlook

In this paper, we conducted an extensive Monte Carlo study including some sensitivity analyses to compare the performance of different Bayesian choice models representing between-segment and/or within-segment consumer heterogeneity. Summing up, the core finding from our simulation study is that the HB-MNL appears to be highly robust against violations in its assumption of a single normal distribution of consumer preferences. The MoN-MNL and the DPM-MNL model on the other hand overestimate the "true" number of components in many cases, which led to a kind of overfitting and as a result of that to large absolute errors regarding parameter recovery and prediction accuracy (independent of which model selection procedure was applied to find the best MoN-MNL models). The latter was particularly distinctive for less complex treatments and for data sets with a low inner-segment heterogeneity. The LC-MNL model proved to be the definitely best approach to recover the "true" number of segments (78%), especially for symmetric treatments concerning segment sizes. The MoN-MNL and DPM-MNL models clearly failed with regard to this criterion, even if the "shut down" procedure suggested by Rossi (2014) provided a much better recovery rate (38%) compared to other model selection procedures. This is especially noteworthy since beyond parameter recovery and prediction accuracy the identification of "true" segment structures is of particular importance for managers who usually do not know the "true" number of segments. Surprisingly, the HB-MNL model performed significantly better or at least as good as all other models as far as parameter recovery (the identification of "true" utility structures) and prediction accuracy is concerned. Regarding model fit, which we consider as not such important for practical applications, only DPM-MNL and MoN-MNL models performed slightly better due to their higher flexibility. Note that some of these findings are also in line with the empirical results reported in Voleti et al. (2017), who especially emphasized the good performance of the HB-MNL model for predictive purposes. Even so, the authors found out that the DPM-MNL model outperformed

[17] The upper part of Fig. 5 in the Appendix additionally displays the fitted population-level distributions obtained for the MoN-MNL and HB-MNL models and demonstrates that the MON-MNL models struggle to recover the true segment structures adequately even for the treatments with a large separation of segments (left upper and lower panels). In contrast, the HB-MNL models fail to re-estimate existing segment structures at the upper level by definition due to its assumption of a single normal distribution. For the sake of completeness and as contrast, the corresponding posterior means distributions again show up in the lower part of Fig. 5. Therefore, in the light of this finding, it may be risky to rely on the fitted population distributions for related marketing decisions, in particular as true segment structures are unknown in empirical applications. See Sect. 4 for a further discussion on this issue.

all other models in their empirical study. Regarding the choice of the model, the HB-MNL model comes off as the clear winner of our Monte Carlo study.

From our perspective, the findings of our Monte Carlo study provide the following managerial implications: (1) Parameter recovery and predictive accuracy are very important criteria (as opposed to model fit) for managerial decision-making, as outlined in the introduction. Since the HB-MNL model either outperformed all other models or at least performed on eye-level with them with regard to all performance measures used, it apparently represents a highly robust choice model under diverse conditions including multimodal preference structures. It did also not fail under very specific conditions we investigated using interaction analyses, but on the contrary performed particularly well in the case of a high number of attributes and attribute levels (i.e. a large number of parameters) with respect to absolute recovery and prediction errors. In addition, DPM-MNL and MoN-MNL models provided huge prediction errors for deterministic utilities in treatments with a large inner-segment heterogeneity, a factor which is not observable in empirical data prior to model estimation. Since MoN-MNL and DPM-MNL models are much more complex (including the need to specify a much larger number of prior settings) and standard software is not available to date, we can recommend practitioners to continue using the well-established HB-MNL model for market (preference) simulations. Note that we ran 200,000 burn-in iterations for each model to ensure convergence of the markov chains, which is of course essential in practical applications, too. (2) If managers work on a segment perspective to design products and related marketing activities, the LC-MNL model can definitely be recommended to identify true segment structures as far as such exist. In contrast, the ability of both DPM-MNL and MoN-MNL models to recover the "true" number of segments was considerably worse and rather disappointing in our study. One possibility for managers to nevertheless address inner-segment heterogeneity would be to estimate a LC-MNL model at first and subsequently a HB-MNL model for (some of) the identified segments. Alternatively, the segment-specific part-worths obtained from the LC-MNL model could be weighted by a respondent's posterior segment membership probabilities to arrive at individual part-worth utility estimates for each individual (e.g., Vermunt and Magidson 2007). (3) Of course, findings on predictive validity from artificial data sets need not coincide with those from empirical settings with real data. Synthetic data may contain inadvertent biases from not considering real-world phenomena like simplification strategies of respondents or respondent fatigue in later choice tasks (e.g., Selka et al. 2014). Note that we considered the latter issue with an experimental factor that limited the number of choice tasks to a manageable number following related meta studies. In a recent empirical study for eleven CBC data sets, Voleti et al. (2017) reported higher hit rates and higher hit probabilties for the DPM-MNL model compared to the HB-MNL, MoN-MNL, and LC-MNL models. The DPM-MNL model improved hit rates / hit probabilities on average by 5% / 3% over the HB-MNL, however the HB-MNL outperformed the MoN-MNL by 2% / 8% and LC-MNL models by 3% / 9%, on average. In our study, holdout sample hit rates were comparable across the four models, whereas the HB-MNL (DPM-MNL) performed clearly best (worst) in predicting deterministic utilities. More research is needed here to explore the differences in predictive accuracy between the DPM-MNL and the HB-MNL in empirical versus artificial settings. Nevertheless, the HB-MNL also predicted surprisingly well in the empirical

study of Voleti et al. (2017). We elaborate on this issue still in more detail below in our outlook on future research opportunities.

Future work should further verify if our findings hold for different distributions of heterogeneity than assumed in the present study. For example, if the distribution of inner-segment heterogeneity is rather skewed, one might expect a superior performance of the MoN-MNL or the DPM-MNL models compared to the HB-MNL, LC-MNL and aggregate MNL models. In a Monte Carlo study, Ebbes et al. (2015) for example additionally estimated so-called DPP models. In DPP models, the distribution of part-worth utilities is drawn from a Dirichlet Process, with the resulting part-worth utilities representing a mixture of discrete vectors. Performance measures for the DPP models did not differ significantly from the measures obtained for the MoN models in the study of Ebbes et al. (2015).[18] However, the authors conjectured that the DPPs will outperform MoN models if the distribution of inner-segment heterogeneity differs from a normal distribution. It should be noted that Andrews et al. (2002a) found no differences in measures of performances between different choice models when comparing normally distributed preferences to gamma distributed preferences. However, they only compared a LC-MNL model, a HB-MNL model and an aggregate model and did not consider the MoN-MNL and the DPM-MNL models. Kim et al. (2004) concluded that the recovery performance of models with a Dirichlet Process prior was getting worse for data sets with a mixture of skewed distributions compared to data sets with a mixture of normal distributions. However, they did not compare the recovery performance to a HB-MNL model with an unimodal distribution of heterogeneity or to LC-MNL models.

Future research could also investigate how well the different models predict truly out-of-sample, i.e. not just for new observations of the respondents in holdout choice sets but for entirely new respondents (e.g., Pachali et al. 2020). In this case, possible concerns that a model is trained not only to fit the data well, but also may favor "overfitting" holdout choices could be eliminated. Basically, there are several ways to predict the choice behavior out-of-sample. One option would be using the posterior means of the respondents' part-worths from the estimation sample and just integrating over this distribution of posterior means for predictions. Alternatively, one might simulate draws from the density of the respondents' posterior means instead of using the posterior means resulting directly from the Markov chain after convergence. In both cases, predictions for the new sample would be based on the posterior means of the respondents of the estimation sample. However, Pachali et al. (2020) showed for the normal distribution that the heterogeneity of respondents would be underestimated from the posterior means of individual part-worths, a finding that we could confirm not only for the HB-MNL model with its single normal distribution but also for the MoN-MNL model with its mixture-of-normals, please directly compare the posterior mean distributions versus the fitted population distributions displayed for selected treatments in Fig. 5 in the Appendix. In order to adequately "exploit" the heterogeneity of respondents, it therefore

[18] Since groups of respondents share identical part-worth utilities, the DPP-MNL model is closely related to the LC-MNL model. As the DPP-MNL model performed very worse and often not better than a chance model concerning hit probabilities in the empirical CBC study of Voleti et al. (2017) we did not include it in our study.

seems more reasonable at first glance to use the fitted population distributions for out-of-sample predictions. In order to get an impression how results could change out-of-sample, we compared the HB-MNL model with the MoN-MNL model for a randomly selected treatment with a large separation and small within-segment heterogeneity (i.e. a clearly multimodal preference structure). For this, we estimated the two models for 400 respondents (estimation sample), threw away the posterior mean estimates of the estimation sample and instead simulated 400 random draws from the fitted population-level distributions,[19] and finally predicted the choices for each of the 200 new respondents (validation sample) based on these 400 simulated draws. Out-of sample hit rates of the two models were highly comparable (HB-MNL: 82.5%; MoN-MNL: 82.3%), indicating that the HB-MNL performs competitive out-of-sample, too. Of course, this represents just one instance and much more research is necessary to generalize this finding. On the other hand, this result is not even surprising with regard to the plots shown in Fig. 5 which already suggested that the MoN-MNL model was not working as expected on the upper level in recovering the existing segment structures. Hence, the Mon-MNL model could not play its theoretical advantage against the HB-MNL model which is expected to predict worse for a new sample of respondents by definition if population-level preferences are (clearly) multimodal. Given that (1) the HB-MNL model has been proven to recover multimodal distributions of individual-level parameters in the estimation sample despite its very wrong assumption of a single normal population distribution, and (2) the MoN-MNL model might be not able to recover a multimodal population distribution satisfactory even if a clear separation of true segments exists (like in our study), for both models the bias from an underestimation of the degree of heterogeneity when using the distribution of posterior means of the estimation sample might be eventually smaller than the bias from using a wrong population distribution.

Finally, future research could analyze the performance of the competing models when taking into account simplification strategies of respondents, which are known to occur in empirical studies. Simplification strategies can, for example, be the result of (a) straightlining behavior of respondents who pay attention to only one or two key attributes when choosing brands, (b) some kind of cheating behavior of professional respondents as can be more and more observed in online panels, or (c) simply boringness of respondents (Hein et al. 2020). Simplification strategies reduce the quality of the data compared to artificial studies and thus may affect the relative performance of the different models studied in this paper. To the best of our knowledge, no simulation study has yet compared the performance of the aggregate MNL, LC-MNL, HB-MNL, MoN-MNL and DPM-MNL models in the presence of simplification strategies of at least parts of respondents. Moreover, including the sample size as an additional experimental factor might provide further insights about the overlapping mixtures problem (Kim et al. 2004), which affects the performance of DPM and MoN models.

Appendix

See Tables 9, 10 and Fig. 5

[19] For the MoN-MNL model, the number of draws for each mixture component was determined by the estimated membership probabilities.

Table 9 F-Tests[a] of main and interaction effects on parameter recovery, goodness-of-fit and predictive accuracy (N = 1200; p-values in parentheses). Note that performance measures were calculated based on 200 individual draws and that only the best LC-MNL and MoN-MNL solutions as provided by the model selection (based on the log marginal likelihood) were included in the ANOVAs

	Parameter recovery			Goodness-of-fit			Predictive accuracy	
Source (degrees of freedom)	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Model (4)	**395.409 (< 0.001)**	**57.560 (< 0.001)**	**4313.357 (< 0.001)**	**2870.015 (< 0.001)**	**4161.578 (< 0.001)**	**3101.658 (< 0.001)**	**499.870 (< 0.001)**	**61.490 (< 0.001)**
Model complexity (2)	**265.126 (< 0.001)**	**119.034 (< 0.001)**	**169.450 (< 0.001)**	**21.757 (< 0.001)**	**417.010 (< 0.001)**	**246.096 (< 0.001)**	**376.681 (< 0.001)**	**29.611 (< 0.001)**
Number of segments (2)	**33.830 (< 0.001)**	1.697 (0.184)	**23.466 (< 0.001)**	**85.211 (< 0.001)**	**90.603 (< 0.001)**	**84.235 (< 0.001)**	**65.084 (< 0.001)**	**4.842 (0.008)**
Separation (1)	**370.546 (< 0.001)**	**23.730 (< 0.001)**	1.869 (0.172)	**283.303 (< 0.001)**	**200.798 (< 0.001)**	**104.863 (< 0.001)**	**276.691 (< 0.001)**	**50.614 (< 0.001)**
Heterogeneity (1)	3.622 (0.057)	**53.288 (< 0.001)**	**212.708 (< 0.001)**	**49.435 (< 0.001)**	**131.604 (< 0.001)**	**128.386 (< 0.001)**	**29.509 (< 0.001)**	**34.437 (< 0.001)**
Segment masses (1)	**20.010 (< 0.001)**	0.488 (0.485)	0.886 (0.347)	1.553 (0.213)	2.410 (0.121)	**17.152 (< 0.001)**	**12.345 (< 0.001)**	2.822 (0.093)
Number of choice sets per respondent (1)	**316.843 (< 0.001)**	**127.873 (< 0.001)**	**35.758 (< 0.001)**	**26.069 (< 0.001)**	**40.431 (< 0.001)**	**16.615 (< 0.001)**	**786.429 (< 0.001)**	**202.803 (< 0.001)**
Model × Model complexity	**28.338 (< 0.001)**	**31.419 (< 0.001)**	**47.445 (< 0.001)**	**43.869 (< 0.001)**	**88.915 (< 0.001)**	**44.139 (< 0.001)**	**67.010 (< 0.001)**	**34.083 (< 0.001)**
Model × Number of segments	**17.724 (< 0.001)**	**3.730 (< 0.001)**	**8.142 (< 0.001)**	**46.221 (< 0.001)**	**36.137 (< 0.001)**	**40.437 (< 0.001)**	**17.944 (< 0.001)**	**4.560 (< 0.001)**
Model × Separation	**175.736 (< 0.001)**	**31.519 (< 0.001)**	**22.570 (< 0.001)**	**401.614 (< 0.001)**	**376.212 (< 0.001)**	**239.372 (< 0.001)**	**191.010 (< 0.001)**	**33.688 (< 0.001)**
Model × Heterogeneity	1.976 (0.096)	**28.439 (< 0.001)**	**89.139 (< 0.001)**	**10.344 (< 0.001)**	**13.737 (< 0.001)**	**7.442 (< 0.001)**	**2.561 (0.037)**	**39.394 (< 0.001)**
Model × Segment masses	**6.873 (< 0.001)**	0.184 (0.947)	0.439 (0.780)	0.668 (0.614)	0.636 (0.637)	**12.293 (< 0.001)**	**3.267 (0.011)**	0.215 (0.930)

Table 9 (continued)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Model × Number of choice sets per respondent	**12.485 (< 0.001)**	1.029 (0.391)	**15.319 (< 0.001)**	**10.303 (< 0.001)**	**39.142 (< 0.001)**	**32.911 (< 0.001)**	**21.609 (< 0.001)**	**3.872 (0.004)**
Model complexity × Number of segments	**3.705 (0.005)**	**2.590 (0.035)**	**2.777 (0.026)**	**4.934 (0.001)**	**5.981 (< 0.001)**	**6.214 (< 0.001)**	**14.765 (< 0.001)**	**3.412 (0.009)**
Model complexity × Separation	**11.829 (< 0.001)**	2.839 (0.059)	**3.649 (0.026)**	**12.271 (< 0.001)**	**8.019 (< 0.001)**	**10.547 (< 0.001)**	**15.362 (< 0.001)**	**3.524 (0.03)**
Model complexity × Heterogeneity	**61.054 (< 0.001)**	**19.203 (< 0.001)**	**5.115 (0.006)**	**51.459 (< 0.001)**	**76.146 (< 0.001)**	**69.268 (< 0.001)**	**9.770 (< 0.001)**	**41.362 (< 0.001)**
Model complexity × Segment masses	**4.060 (0.018)**	1.093 (0.335)	1.701 (0.183)	1.792 (0.167)	**3.517 (0.030)**	2.678 (0.069)	2.685 (0.069)	0.159 (0.853)
Model complexity × Number of choice sets per respondent	**7.398 (0.007)**	1.311 (0.253)	1.992 (0.158)	0.005 (0.945)	0.307 (0.579)	3.197 (0.074)	**25.608 (< 0.001)**	**20.089 (< 0.001)**
Number of segments × Separation	**19.436 (< 0.001)**	0.454 (0.635)	0.891 (0.411)	**7.749 (< 0.001)**	2.175 (0.114)	**4.761 (0.009)**	**8.077 (< 0.001)**	**3.661 (0.026)**
Number of segments × Heterogeneity	1.905 (0.149)	**6.926 (0.001)**	**10.618 (< 0.001)**	2.196 (0.112)	**3.031 (0.049)**	2.793 (0.062)	**3.767 (0.023)**	**6.284 (0.002)**
Number of segments × Segment masses	**6.403 (0.002)**	1.257 (0.285)	0.311 (0.732)	0.052 (0.949)	0.217 (0.805)	**4.192 (0.015)**	**3.832 (0.022)**	0.513 (0.599)
Number of segments × Number of choice sets per respondent	1.988 (0.137)	0.876 (0.417)	0.008 (0.992)	1.709 (0.182)	0.992 (0.371)	0.439 (0.645)	**4.781 (0.009)**	**4.456 (0.012)**
Separation × Heterogeneity	**4.698 (0.030)**	3.292 (0.070)	**4.122 (0.043)**	**8.739 (0.003)**	**7.299 (0.007)**	**4.027 (0.045)**	2.836 (0.092)	2.555 (0.11)
Separation × Segment masses	**10.715 (0.001)**	1.213 (0.271)	1.160 (0.282)	1.466 (0.226)	3.034 (0.082)	0.248 (0.619)	**5.125 (0.024)**	0.026 (0.872)
Separation × Number of choice sets per respondent	**31.213 (< 0.001)**	3.606 (0.058)	0.141 (0.707)	**3.950 (0.047)**	3.312 (0.069)	2.233 (0.135)	**31.851 (< 0.001)**	**9.260 (0.002)**

Table 9 (continued)

Source (degrees of freedom)	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Heterogeneity × Segment masses	**19.767 (< 0.001)**	0.153 (0.695)	0.030 (0.863)	0.032 (0.858)	0.180 (0.671)	0.016 (0.899)	**10.919 (0.001)**	0.381 (0.537)
Heterogeneity × Number of choice sets per respondent	**26.618 (< 0.001)**	0.857 (0.355)	**8.599 (0.003)**	**25.064 (< 0.001)**	**49.625 (< 0.001)**	**42.294 (< 0.001)**	**17.010 (< 0.001)**	0.048 (0.826)
Segment masses × Number of choice sets per respondent	**6.404 (0.012)**	0.003 (0.958)	0.069 (0.793)	0.466 (0.495)	0.296 (0.586)	0.047 (0.828)	3.019 (0.083)	0.537 (0.464)
R^2 (Adjusted R^2)	0.795 (0.782)	0.533 (0.504)	0.944 (0.940)	0.929 (0.924)	0.949 (0.946)	0.932 (0.928)	0.837 (0.827)	0.552 (0.524)

a: Bold values indicate significant differences between factor levels at the 0.05 level.

Table 10 Comparison of results for different model selection procedures for LC-MNL and MoN-MNL models

Factor	Parameter recovery			Goodness-of-fit			Predictive accuracy	
	Correlation	RMSE	%TrueBetas	PC	RLH	IHR	HHR	RMSE(V)
Model (selection by LML)								
(1) DPM-MNL	$0.964^{4,5*}$	$3.091^{2,3,4,5*}$	$0.623^{2,3,4,5*}$	$0.908^{2,3,4*}$	$0.879^{2,3,4*}$	$0.947^{2,3,4*}$	$0.824^{3,4,5*}$	$7.469^{2,3,4,5*}$
(2) HB-MNL	0.968^{4*}	$1.650^{1,3,4,5*}$	$0.746^{1,3,4,5*}$	$0.888^{1,3,4*}$	$0.855^{1,3,4,5*}$	$0.936^{1,3,4*}$	0.832^{4*}	$4.199^{1,3,4,5*}$
(3) LC-MNL	0.967^{4*}	$2.159^{1,2,5*}$	$0.014^{1,2,5*}$	$0.828^{1,2,4,5*}$	$0.785^{1,2,4,5*}$	$0.905^{1,2,4,5*}$	$0.835^{1,4*}$	$5.395^{1,2,4,5*}$
(4) MNL	$0.926^{1,2,3,5*}$	$2.466^{1,2*}$	$0.021^{1,2,5*}$	$0.583^{1,2,3,5*}$	$0.567^{1,2,3,5*}$	$0.787^{1,2,3,5*}$	$0.765^{1,2,3,5*}$	$6.275^{1,2,3*}$
(5) MoN-MNL	$0.971^{1,4*}$	$2.608^{1,2,3*}$	$0.585^{1,2,3,4*}$	$0.903^{3,4*}$	$0.873^{2,3,4*}$	$0.943^{3,4*}$	$0.838^{1,4*}$	$6.352^{1,2,3*}$
Model (selection by WAIC)								
(1) DPM-MNL	$0.964^{4,5*}$	$3.091^{2,3,4,5*}$	$0.623^{2,3,4*}$	$0.908^{2,3,4*}$	$0.879^{2,3,4*}$	$0.947^{2,3,4*}$	$0.824^{4,5*}$	$7.469^{2,3,4,5*}$
(2) HB-MNL	0.968^{4*}	$1.650^{1,3,4,5*}$	$0.746^{1,3,4,5*}$	$0.888^{1,3,4*}$	$0.855^{1,3,4*}$	$0.936^{1,3,4*}$	0.832^{4*}	$4.199^{1,3,4,5*}$
(3) LC-MNL	0.967^{4*}	$2.168^{1,2*}$	$0.013^{1,2,5*}$	$0.827^{1,2,4,5*}$	$0.784^{1,2,4,5*}$	$0.904^{1,2,4,5*}$	0.834^{4*}	$5.410^{1,2,4*}$
(4) MNL	$0.926^{1,2,3,5*}$	$2.466^{1,2*}$	$0.021^{1,2,5*}$	$0.583^{1,2,3,5*}$	$0.567^{1,2,3,5*}$	$0.787^{1,2,3,5*}$	$0.765^{1,2,3,5*}$	$6.275^{1,2,3*}$
(5) MoN-MNL	$0.973^{1,4*}$	$2.445^{1,2*}$	$0.616^{2,3,4*}$	$0.900^{3,4*}$	$0.869^{3,4*}$	$0.942^{3,4*}$	$0.841^{1,4*}$	$5.951^{1,2*}$
Model (selection by DIC)								
(1) DPM-MNL	$0.964^{4,5*}$	$3.091^{2,3,4,5*}$	$0.623^{2,3,4,5*}$	$0.908^{2,3,4*}$	$0.879^{2,3,4*}$	$0.947^{2,3,4*}$	$0.824^{4,5*}$	$7.469^{2,3,4,5*}$
(2) HB-MNL	0.968^{4*}	$1.650^{1,3,4,5*}$	$0.746^{1,3,4,5*}$	$0.888^{1,3,4*}$	$0.855^{1,3,4*}$	$0.936^{1,3,4*}$	0.832^{4*}	$4.199^{1,3,4,5*}$
(3) LC-MNL	0.966^{4*}	$2.167^{1,2*}$	$0.014^{1,2,5*}$	$0.824^{1,2,4,5*}$	$0.781^{1,2,4,5*}$	$0.903^{1,2,4,5*}$	0.833^{4*}	$5.420^{1,2,4*}$
(4) MNL	$0.926^{1,2,3,5*}$	$2.466^{1,2*}$	$0.021^{1,2,5*}$	$0.583^{1,2,3,5*}$	$0.567^{1,2,3,5*}$	$0.787^{1,2,3,5*}$	$0.765^{1,2,3,5*}$	$6.275^{1,2,3*}$
(5) MoN-MNL	$0.971^{1,4*}$	$2.532^{1,2*}$	$0.588^{1,2,3,4*}$	$0.901^{3,4*}$	$0.871^{3,4*}$	$0.942^{3,4*}$	$0.839^{1,4*}$	$6.222^{1,2*}$
Model (selection by shut down)								
(1) DPM-MNL	0.964^{4*}	$3.091^{2,3,4,5*}$	$0.623^{2,3,4,5*}$	$0.908^{2,3,4*}$	$0.879^{2,3,4*}$	$0.947^{2,3,4*}$	$0.824^{4,5*}$	$7.469^{2,3,4,5*}$
(2) HB-MNL	0.968^{4*}	$1.650^{1,3,4,5*}$	$0.746^{1,3,4,5*}$	$0.888^{1,3,4*}$	$0.855^{1,3,4,5*}$	$0.936^{1,3,4*}$	0.832^{4*}	$4.199^{1,3,4,5*}$
(3) LC-MNL	0.966^{4*}	$2.216^{1,2,5*}$	$0.016^{1,2,5*}$	$0.835^{1,2,4,5*}$	$0.792^{1,2,4,5*}$	$0.909^{1,2,4,5*}$	0.832^{4*}	$5.550^{1,2,5*}$
(4) MNL	$0.926^{1,2,3,5*}$	$2.466^{1,2*}$	$0.021^{1,2,5*}$	$0.583^{1,2,3,5*}$	$0.567^{1,2,3,5*}$	$0.787^{1,2,3,5*}$	$0.765^{1,2,3,5*}$	$6.275^{1,2*}$
(5) MoN-MNL	0.970^{4*}	$2.630^{1,2,3*}$	$0.572^{1,2,3,4*}$	$0.903^{3,4*}$	$0.873^{2,3,4*}$	$0.943^{3,4*}$	$0.836^{1,4*}$	$6.457^{1,2,3*}$

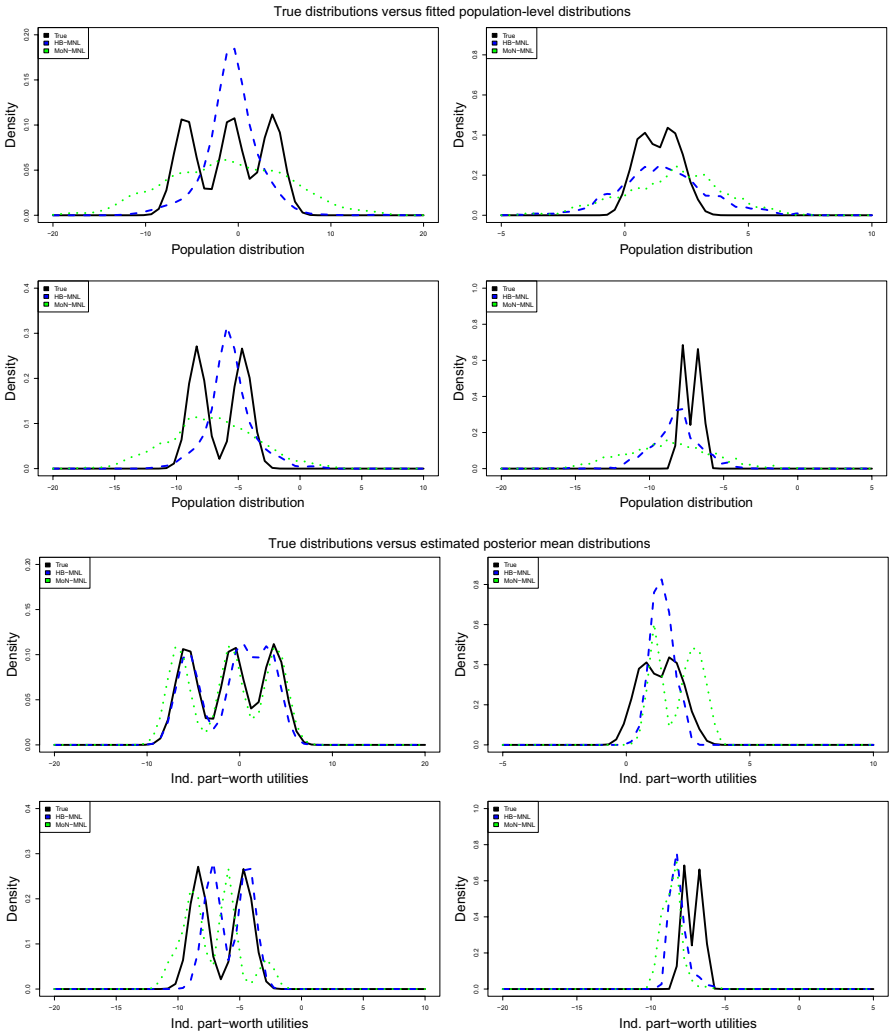

Fig. 5 Selected density plots for "true" distributions of part-worth utilities (black lines) versus re-estimated distributions of part-worth utilities by model type (HB-MNL: blue lines; MoN-MNL: green lines). Upper left panel: 3 "true" segments, large separation, small heterogeneity. Upper right panel: 2 segments, small separation, large heterogeneity. Lower left panel: 2 segments, large separation, large heterogeneity. Lower right panel: 2 segments, small separation, small heterogeneity. The upper panel displays fitted population-level distributions, the lower panel displays the distributions of posterior means of part-worth utilities

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All authors declare that they have no conflicts of interest.

References

Allenby GM, Ginter JL (1995) Using extremes to design products and segment markets. J Mark Res 32:392–403. https://doi.org/10.1177/002224379503200402

Allenby GM, Rossi PE (1998) Marketing models of consumer heterogeneity. Journal of Econometrics 89:57–78. https://doi.org/10.1016/S0304-4076(98)00055-4

Allenby GM, Arora N, Ginter JL (1995) Incorporating prior knowledge into the analysis of conjoint studies. J Mark Res 32:152–162. https://doi.org/10.1177/002224379503200203

Allenby GM, Arora N, Ginter JL (1998) On the heterogeneity of demand. J Mark Res 35:384–389. https://doi.org/10.1177/002224379803500308

Andrews RL, Currim IS (2003) A comparison of segment retention criteria for finite mixture logit models. J Mark Res 40:235–243. https://doi.org/10.1509/jmkr.40.2.235.19225

Andrews RL, Ainslie A, Currim IS (2002a) An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. J Mark Res 39:479–487. https://doi.org/10.1509/jmkr.39.4.479.19124

Andrews RL, Ansari A, Currim IS (2002b) Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. J Mark Res 39:87–98. https://doi.org/10.1509/jmkr.39.1.87.18936

Andrews RL, Ainslie A, Currim IS (2008) On the recoverability of choice behaviors with random coefficients choice models in the context of limited data and unobserved effects. Manage Sci 54:83–99. https://doi.org/10.1287/mnsc.1070.0749

Ansari A, Mela CF (2003) E-Customization. J Mark Res 40:131–145. https://doi.org/10.1509/jmkr.40.2.131.19224

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174. https://doi.org/10.1214/aos/1176342871

Burda M, Harding M, Hausman J (2008) A Bayesian mixed logit–probit model for multinomial choice. Journal of Econometrics 147:232–246. https://doi.org/10.1016/j.jeconom.2008.09.029

Chen Y, Iyengar R, Iyengar G (2017) Modeling multimodal continuous heterogeneity in conjoint analysis - a sparse learning approach. Mark Sci 36:140–156. https://doi.org/10.1287/mksc.2016.0992

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, New York

Conley TG, Hansen CB, McCulloch RE, Rossi PE (2008) A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics 144:276–305. https://doi.org/10.1016/j.jeconom.2008.01.007

Danaf M, Becker F, Song X, Atasoy B, Ben-Akiva M (2019) Online discrete choice models: applications in personalized recommendations. Decis Support Syst 119:35–45. https://doi.org/10.1016/j.dss.2019.02.003

DeSarbo WS, Ramaswamy V, Cohen SH (1995) Market segmentation with choice-based conjoint analysis. Mark Lett 6:137–147. https://doi.org/10.1007/BF00994929

Dias JG, Vermunt JK (2007) Latent class modeling of website users' search patterns: Implications for online market segmentation. J Retail Consum Serv 14:359–368. https://doi.org/10.1016/j.jretconser.2007.02.007

Ebbes P, Liechty JC, Grewal R (2015) Attribute-level heterogeneity. Manage Sci 61:885–897. https://doi.org/10.1287/mnsc.2014.1898

Elshiewy O, Guhl D, Boztuğ Y (2017) Multinomial logit models in marketing-from fundamentals to state-of-the-art. Marketing ZFP 39:32–49. https://doi.org/10.15358/0344-1369-2017-3-32
Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:557–588. https://doi.org/10.1080/01621459.1995.10476550
Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230. https://doi.org/10.1214/aos/1176342360
Frühwirth-Schnatter S (2004) Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. Econom J 7:143–167. https://doi.org/10.1111/j.1368-423X.2004.00125.x
Frühwirth-Schnatter S (2006) Finite mixture and markov switching models. Springer Science & Business Media, New York
Frühwirth-Schnatter S, Tüchler R, Otter T (2004) Bayesian analysis of the heterogeneity model. J Bus Econ Stat 22:2–15. https://doi.org/10.1198/073500103288619331
Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–472. https://doi.org/10.1214/ss/1177011136
Goeken N, Kurz P, Steiner WJ (2021) Hierarchical Bayes conjoint choice models - Model framework, Bayesian inference, model selection, and interpretation of estimation results. Marketing ZFP 43:49–66. https://doi.org/10.15358/0344-1369-2021-3-49
Hauser JR (1978) Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach. Oper Res 26:406–421. https://doi.org/10.1287/opre.26.3.406
Hauser JR, Rao VR (2004) Conjoint analysis, related modeling, and applications. In: Wind Y, Green PE (eds) Marketing research and modeling: Progress and prospects. Springer Science & Business Media, New York, pp 141–168
Hein M, Kurz P, Steiner WJ (2019) On the effect of HB covariance matrix prior settings: A simulation study. Journal of Choice Modelling 31:51–72. https://doi.org/10.1016/j.jocm.2019.02.001
Hein M, Kurz P, Steiner WJ (2020) Analyzing the capabilities of the HB logit model for choice-based conjoint analysis: A simulation study. J Bus Econ 90:1–36. https://doi.org/10.1007/s11573-019-00927-4
Hoogerbrugge M, van der Wagt K (2006) How many choice tasks should we ask? Proceedings of the 2006 Sawtooth Software Conference 97–110
Horowitz JL, Nesheim L (2021) Using penalized likelihood to select parameters in a random coefficients multinomial logit model. Journal of Econometrics 222:44–55. https://doi.org/10.1016/j.jeconom.2019.11.008
Jervis SM, Lopetcharat K, Drake MA (2012) Application of ethnography and conjoint analysis to determine key consumer attributes for latte-style coffee beverages. J Sens Stud 27:48–58. https://doi.org/10.1111/j.1745-459X.2011.00366.x
Kamakura WA, Russell GJ (1989) A probabilistic choice model for market segmentation and elasticity structure. J Mark Res 26:379–390. https://doi.org/10.1177/002224378902600401
Kamakura WA, Wedel M, Agrawal J (1994) Concomitant variable latent class models for conjoint analysis. Int J Res Mark 11:451–464. https://doi.org/10.1016/0167-8116(94)00004-2
Keane MP, Ketcham J, Kuminoff N, Neal T (2021) Evaluating consumers' choices of Medicare Part D plans: A study in behavioral welfare economics. Journal of Econometrics 222:107–140. https://doi.org/10.1016/j.jeconom.2020.07.029
Kim JG, Menzefricke U, Feinberg FM (2004) Assessing heterogeneity in discrete choice models using a Dirichlet process prior. Rev Mark Sci 2:1–39. https://doi.org/10.2202/1546-5616.1003
Kim J, Allenby GM, Rossi PE (2007) Product attributes and models of multiple discreteness. Journal of Econometrics 138:208–230. https://doi.org/10.1016/j.jeconom.2006.05.020
Kuhfeld WF (2019) Orthogonal arrays. SAS Institute Inc, Technical Paper
Kurz P, Binner S (2012) The individual choice task threshold: Need for variable number of choice tasks. Proceedings of the 2012 Sawtooth Software Conference 111–127
Lenk PJ, DeSarbo WS (2000) Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika 65:93–119. https://doi.org/10.1007/BF02294188
Lenk PJ, DeSarbo WS, Green PE, Young MR (1996) Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. Mark Sci 15:173–191. https://doi.org/10.1287/mksc.15.2.173
Li Y, Ansari A (2014) A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models. Manage Sci 60:1161–1179. https://doi.org/10.1287/mnsc.2013.1811

Louviere JJ, Woodworth G (1983) Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. J Mark Res 20:350–367. https://doi.org/10.1177/002224378302000403

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

Moore WL, Gray-Lee J, Louviere JJ (1998) A cross-validity comparison of conjoint analysis and choice models at different levels of aggregation. Mark Lett 9:195–207. https://doi.org/10.1023/A:1007913100332

Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. J Roy Stat Soc Ser B (methodol) 56:3–48

Ogawa K (1987) An approach to simultaneous estimation and segmentation in conjoint analysis. Mark Sci 6:66–81. https://doi.org/10.1287/mksc.6.1.66

Otter T, Tüchler R, Frühwirth-Schnatter S (2004) Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. Int J Res Mark 21:285–297. https://doi.org/10.1016/j.ijresmar.2003.11.002

Pachali MJ, Kurz P, Otter T (2020) How to generalize from hierarchical model? Quant Mark Econ 18:343–380. https://doi.org/10.1007/s11129-020-09226-7

Paetz F, Hein M, Kurz P, Steiner WJ (2019) Latent class conjoint choice models: A guide for model selection, estimation, validation, and interpretation of results. Marketing ZFP 41:3–20. https://doi.org/10.15358/0344-1369-2019-4-3

Paetz F, Steiner WJ (2017) The benefits of incorporating utility dependencies in finite mixture probit models. OR Spectrum 39:793–819. https://doi.org/10.1007/s00291-017-0478-y

R Core Team (2020) R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria. URL https://www.R-project.org/

Rodríguez CE, Walker SG (2014) Label switching in Bayesian mixture models: Deterministic relabeling strategies. J Comput Graph Stat 23:25–45. https://doi.org/10.1080/10618600.2012.735624

Rossi PE (2014) Bayesian non- and semi-parametric methods and applications. Princeton University Press, Princeton

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. Mark Sci 15:321–340. https://doi.org/10.1287/mksc.15.4.321

Rossi PE, Allenby GM, McCulloch R (2005) Bayesian statistics and marketing. John Wiley & Sons, Chichester

Rossi PE (2019) bayesm: Bayesian inference for marketing/micro-econometrics. R package version 3.1–4. https://CRAN.R-project.org/package=bayesm

Selka S, Baier D, Kurz P (2014) The validity of conjoint analysis: An investigation of commercial studies over time. In: Spiliopoulou M, Schmidt-Thieme L, Janning R (eds) Data analysis, machine learning and knowledge discovery. Springer, Cham, pp 227–234. https://doi.org/10.1007/978-3-319-01595-8_25

Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sin 4:639–650

Sawtooth Software (2016) Software for hierarchical Bayes estimation for CBC data, CBC/HB v5. Sawtooth Software, Inc

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J Roy Stat Soc: Ser B (methodol) 64:583–639. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2014) The deviance information criterion: 12 years on. J Roy Stat Soc: Ser B (methodol) 76:485–493. https://doi.org/10.1111/rssb.12062

Street DJ, Burgess L (2007) The construction of optimal stated choice experiments: Theory and methods. John Wiley & Sons, New Jersey

Street DJ, Burgess L, Louviere JJ (2005) Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. Int J Res Mark 22:459–470. https://doi.org/10.1016/j.ijresmar.2005.09.003

Train KE (2009) Discrete choice methods with simulation, 2nd edn. Cambridge University Press, New York

Tuma M, Decker R (2013) Finite mixture models in market segmentation: A review and suggestions for best practices. Electronic Journal of Business Research Methods 11:2–15

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Verlegh PWJ, Schifferstein HNJ, Wittink DR (2002) Range and number-of-levels effects in derived and stated measures of attribute importance. Mark Lett 13:41–52. https://doi.org/10.1023/A:1015063125062

Vermunt JK, Magidson J (2007) Latent class analysis with sampling weights: A maximum-likelihood approach. Sociological Methods and Research 36:87–111. https://doi.org/10.1177/0049124107301965

Voleti S, Srinivasan V, Ghosh P (2017) An approach to improve the predictive power of choice-based conjoint analysis. Int J Res Mark 34:325–335. https://doi.org/10.1016/j.ijresmar.2016.08.007

Vriens M, Wedel M, Wilms T (1996) Metric conjoint segmentation methods: A Monte Carlo comparison. J Mark Res 33:73–85. https://doi.org/10.1177/002224379603300107

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594

Webb R, Mehta N, Levy I (2021) Assessing consumer demand with noisy neural measurements. Journal of Econometrics 222:89–106. https://doi.org/10.1016/j.jeconom.2020.07.028

Wedel M, Kamakura WA, Arora N, Bemmaor A, Chiang J, Elrod T, Johnson R, Lenk PJ, Neslin S, Poulsen CS (1999) Discrete and continuous representations of unobserved heterogeneity in choice modeling. Mark Lett 10:219–232. https://doi.org/10.1023/A:1008054316179

Wirth R (2010) HB-CBC, HB-Best-Worst-CBC or no HB at all? Proceedings of the 2010 Sawtooth Software Conference 321–355

Zhao L, Shi J, Shearon TH, Li Y (2015) A Dirichlet process mixture model for survival outcome data: Assessing nationwide kidney transplant centers. Stat Med 34:1404–1416. https://doi.org/10.1002/sim.6438