Research Articles

Hierarchical Bayes Conjoint Choice Models – Model Framework, Bayesian Inference, Model Selection, and Interpretation of Estimation Results

By Nils Goeken*, Peter Kurz, and Winfried J. Steiner

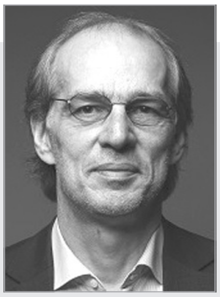
The use of hierarchical Bayes (HB) multinomial logit (MNL) models for measuring consumer preferences is state-of-the-art in choice-based conjoint (CBC) analysis. Here, academic researchers and practitioners mostly utilize by default the normal prior for the first level of the hierarchical model. However, a mixture of normal distributions also appears promising, providing more flexibility to accommodate multimodal preference structures or skewed preference distributions. There are currently two prominent HB-CBC modelling approaches embedding the mixture-of-normals approach: the more widespread mixture-of-normals (MoN)-HB-MNL model, and the Dirichlet process mixture (DPM)-HB-MNL model. In this article, we review the standard HB-MNL (with its normal prior), the MoN-HB-MNL, and the DPM-HB-MNL models, applying them to an empirical multi-country CBC data set. We discuss related Bayesian estimation processes, including model selection issues; compare the statistical performance of the three models in terms of fit and prediction in an empirical study; and show how estimation results can be interpreted.

1. Introduction

Companies and market research institutes often collect and analyse consumer preference data in an attempt to predict market demand for new or modified products, to improve pricing decisions, and/or to segment markets. Currently, the most widely used marketing tool to measure consumer preferences is choice-based conjoint (CBC) analysis (Louviere and Woodworth 1983). Here, preferences for attributes and attribute levels are collected through experimental choice decisions, which closely resembles the choice behaviour of consumers in real marketplaces. The popularity of CBC has grown even more since the introduction of hierarchical Bayesian (HB) estimation techniques (Allenby et al. 1995; Allenby and Ginter 1995; Lenk et al. 1996) that accommodate individual consumer heterogeneity in choice data, and which have become state-of-the-art in marketing theory and practice (e.g. Aribarg et al. 2017; Baumgartner and Steiner 2007; Hein et al. 2019, 2020; Voleti et al. 2017). There is strong empirical evidence that addressing individual preference heterogeneity in CBC studies using HB modelling pays off in terms of statistical model performance (e.g. providing a higher forecasting accuracy) compared to focusing only on a finite number of discrete support points assuming perfectly homogeneous seg-

Nils Goeken is Assistant Professor of Marketing (Akademischer Rat) at Clausthal University of Technology, Julius-Albert-Str. 2, D-38678 Clausthal-Zellerfeld, Germany, Phone: +49/5323 727660, Fax: +49/5323 727659, E-Mail: nils.goeken@tu-clausthal.de.
* Corresponding Author.

Peter Kurz is Managing Partner, Innovation & Methods, at bms marketing research & strategy, Landsberger Strasse 487, D-81241 München, Germany, Phone: +49/89 88969442 Fax: +49/89 88969444, E-Mail: p.kurz@bms-net.de.



Winfried Steiner is Professor of Marketing at Clausthal University of Technology, Julius-Albert-Str. 2, D-38678 Clausthal-Zellerfeld, Germany, Phone: +49/5323 727650, Fax: +49/5323 72 7659, E-Mail: winfried.steiner@tu-clausthal.de.

ments (e.g. Allenby et al. 1998; Andrews et al. 2002a; Elshiewy et al. 2017; Moore 2004; Natter and Feurstein 2002; Voleti et al. 2017). For this reason, we will focus on HB models for CBC data (also known as a continuous modelling approach) in this paper [1]. While discrete modelling approaches may be criticized for oversimplifying the concept of heterogeneity, continuous modelling approaches depend on additional assumptions about the distribution of preference heterogeneity at the population level. In particular, if a unimodal preference distribution is assumed (as is the standard assumption, especially in practical/commercial CBC applications), it can be expected that the continuous approach might not be flexible enough to reproduce existing preference heterogeneity. More advanced HB modelling approaches were (recently) introduced into the marketing literature, which importantly relax this supposed inflexibility of the standard HB model. The aim of this contribution is to provide an overview of the established HB conjoint choice models (including their model framework), Bayesian inference, model selection issues, and interpretation of estimation results. We further apply and compare the different HB approaches in an empirical study for a multi-country CBC data set.

The most widely used model with a continuous representation of heterogeneity for CBC data is the so-called HB-multinomial logit (MNL) model, in the following referred to as “HB-MNL” (Allenby et al. 1995; Allenby and Ginter 1995; Elshiewy et al. 2017; Lenk et al. 1996). Here, using a normal distribution (and therefore a unimodal distribution) has become the standard procedure in the marketing literature to represent preference heterogeneity. However, it is just as likely (or there is no good reason to rule out a priori) that a true distribution of consumer preferences is multimodal. A first indication that the HB-MNL might also adapt well to multimodal distributions despite its unimodal normal prior heterogeneity distribution was provided by Andrews et al. (2002a) from a Monte Carlo study for logit models applied in a scanner data context. The thin tails of the normal distribution nevertheless suggest that the HB-MNL should not be the “go-to” approach to approximate multimodal preference distributions, because individual preference patterns lying at the tails of the normal distribution (i.e. that do not fit well with the assumption of a unimodal distribution) tend to be shrunk to the population mean. This shrinkage, especially in multimodal data settings, could mask important information (e.g. new or different structures in the data) (Rossi et al. 2005, p. 142). In addition, heterogeneity distributions can be highly skewed (e.g. for price coefficients). Hence, preference structures of consumers may be too complex to approximate with one unimodal distribution. The mixture-of-normals HB-MNL model, in the following referred to as “MoN-HB-MNL”, avoids this limited flexibility of the most simple continuous approach of assuming a unimodal prior heterogeneity distribution (Lenk and DeSarbo 2000). Here, the logit model is used to analyse individual preferences at the lower level of the hierarchical model, while a mixture of several multivariate normal distributions is specified as prior heterogeneity distribution for consumers at the population or upper level (Allenby et al. 1998). In doing this, MoN-HB-MNL models can accommodate multimodal, heavy-tailed, and skewed distributions by using a sufficient number of normal components. The model in turn is able to discover new or different structures in data. From a marketing perspective, the MoN-HB-MNL is able to account for more than one segment or component (as implicitly assumed in the simple HB-MNL by using a single normal prior), and therefore allows for both inner-segment and between-segment (preference) heterogeneity. Allenby et al. (1998) and Baumgartner and Steiner (2007) for example found strong support for the application of MoN-HB-MNL models. In their empirical studies, the MoN-HB-MNL model outperformed the HB-MNL model in terms of goodness-of-fit and predictive accuracy. However, comparable to latent class approaches, the number of components must be fixed prior to model estimation in the MoN-HB-MNL approach, and model selection procedures have to be applied to find the “optimal” number of components.

Recently, the Dirichlet process mixture HB-MNL model, referred to as “DPM-HB-MNL” in the following, has been applied to CBC data to capture preference heterogeneity (Voleti et al. 2017). The DPM-HB-MNL model is also able to approximate multimodal, heavy-tailed and/or skewed preference distributions, and is even more flexible than the MoN-HB-MNL model. The part-worth utilities are drawn from continuous distributions (here again, a mixture of multivariate normal distributions), where population means and covariances follow a Dirichlet process. In other words, the continuous distributions are centered around the discrete part-worth utilities of a Dirichlet process (Voleti et al. 2017). A further advantage of the DPM-HB-MNL model beyond its strong flexibility (Krueger et al. 2018) is that the number and composition of the components adjusts as a result a posteriori (Rossi 2014, p. 72). The DPM-HB-MNL can be considered an extension of the MoN-HB-MNL, where the number of components becomes part of the Bayesian estimation process. In an empirical study based on eleven CBC data sets with different characteristics, Voleti et al. (2017) reported a superior predictive validity (measured by out-of-sample hit rates) over several other choice models with discrete (e.g. latent class MNL models) or continuous representations of consumer heterogeneity (e.g. HB-MNL or MoN-HB-MNL models).

In this paper, we compare the performance of HB choice models with the different prior distributions mentioned above by means of a real-life, multi-country CBC data set for tires. We specifically assess the performance of the HB-MNL, the MoN-HB-MNL, and the DPM-HB-MNL models in terms of goodness-of-fit and predictive accuracy. To account for cross-country heterogeneity and analyse the effects of including additional consumer background characteristics, we estimate and compare the three types of models, with versus without these concomitant covariates. In Section 2, we first review the three hierarchical choice models, discuss the related Bayesian estimation processes, and propose measures to assess the statistical performance. In Section 3, we apply and compare the choice models in an empirical study while also addressing model selection issues for the MoN-HB-MNL (in the sense of determining the best solution). We then highlight the need to account for heterogeneity, show the effects of including individual consumer characteristics (concomitant covariates), and provide an example of how to analyse differences in respondents' preferences between countries.

2. Random utility models

All of the choice models presented in the introduction belong to the class of random utility models (RUMs) and are based on the assumption of utility-maximizing consumers. In RUMs, a respondent is assumed to choose the alternative with the highest utility (Silberhorn et al. 2008; Train 2009, p. 14). For our purposes, we will denote the utility that respondent n obtains from alternative j in choice situation s as $U_{njs} = V_{njs} + \varepsilon_{njs}$. V_{njs} represents the deterministic part of the utility, whereas ε_{njs} represents the stochastic part of the utility. We specify the deterministic utility V_{njs} as linear in parameters: $V_{njs} = \beta_n' x_{njs}$. Here, β_n represents the vector of part-worth utilities for respondent n, where the part-worth utilities refer to attributes or attribute levels considered relevant for the choice of alternatives. x_{njs} is a vector for the attribute levels (either coded using dummies or linearly) of alternative j offered to respondent n in choice situation s. β_n can vary according to the heterogeneity distribution (normal distribution vs. mixture of normal distributions) underlying the respective modelling approach. The estimation of all models is fully Bayesian.

2.1. Hierarchical MNL and MoN-HB-MNL models

By assuming that the unknown part ε_{njs} of the utility U_{njs} follows a Gumbel distribution, we obtain the MNL model with closed-form expressions for choice probabilities (Train 2009, p. 34):

$$P_{njs}^{MNL} = \frac{e^{V_{njs}}}{\sum_h e^{V_{nhs}}}. \tag{1}$$

Using a finite mixture of normals, we are able to approximate multimodal heterogeneity structures as well as skewed distributions and distributions with thick tails. The general form of a finite mixture of normals consists of a mixing distribution p_m and the multivariate normal density $\phi(\beta | b_m, W_m)$. The mixing distribution p_m puts mass on M different values of b_m and $W_m, m \in \{1, ..., M\}$:

$$f(\beta | p, \{b_m, W_m\}) = \sum_m p_m \ \phi(\beta | b_m, W_m). \tag{2}$$

Using a sufficient number of components, this approach can approximate any multivariate density (Train 2009, pp. 141–143).

We specify the following hierarchical model for Bayesian inference (Rossi 2014, p. 156; Rossi et al. 2005, pp. 144–145):

$$\begin{array}{ll} \beta_n & = \Delta' z_n + \xi_n \\ \xi_n & \sim \mathcal{N}(b_{l_n}, W_{l_n}), \\ l_n & \sim MN_M(p), \\ \text{vec}(\Delta) & \sim \mathcal{N}(\overline{\delta}, A_\delta^{-1}), \\ p & \sim \text{Dirichlet}(\alpha), \\ b_m & \sim \mathcal{N}(\overline{b}, w^{-1} W_m), \\ W_m & \sim IW(k, \Sigma). \end{array} \tag{3}$$

The $n_z z$-variables represent observable individual background characteristics of each of the $n = 1,...,N$ respondents (also referred to as concomitant variables) and shift the mean of the normal mixture on the basis of these observations. Here, $\Delta \in (n_z,d)$ with d as the dimension of the data (number of part-worth utilities) describes how the means of the part-worth utilities vary as a function of the z-variables. Δ is normally distributed with mean vector δ and covariance matrix A_δ^{-1}. The indicator variable $l_n \in \{1,...,M\}$ represents the outcome of a multinomial distribution and indicates from which component the part-worth utilities of respondent n are drawn. $p \in \mathbb{R}^M$ are the associated probabilities following a Dirichlet distribution. $\alpha \in \mathbb{R}^M$ can be interpreted as a tightness parameter, which has an influence on the masses of the components. Rossi (2014, pp. 72–76) shows that a large α leads to a substantially larger number of components the Dirichlet distribution places mass on. The corresponding population means b_m and the covariance matrices W_m with $m \in \{1,...,M\}$ are normally distributed (with mean vector b and covariance matrix $w^{-1}W_m$), and inverse Wishart distributed (with k degrees of freedom and scale matrix Σ). The dimensions of b_m and W_m depend on the number of parameters (part-worth utilities) to be estimated. Following Allenby et al. (1998) the MoN-HB-MNL model as well as the simple HB-MNL model can be considered special cases of a finite mixture of normals framework. For $M = 1$ we obtain the HB-MNL, for $M > 1$ the MoN-HB-MNL.

2.2. DPM-HB-MNL model

A more recent approach which also accounts for continuous consumer heterogeneity is the DPM-HB-MNL model. As mentioned in the introduction, the DPM-HB-MNL can be considered an extension of the MoN-HB-MNL approach where the number of components become part of the Bayesian estimation framework. This Bayesian estimation framework allows for a countably infinite number of components by placing additional priors on the component parameters. Rossi (2014, p. 59) for example comments on a more flexible approximation of multimodal distributions when a larger number of multivariate components are considered. To obtain the DPM-HB-MNL model, the Dirichlet prior in the MoN approach is replaced by a Dirichlet process:

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

In this setting, the normal distribution (as seen above) accounts for within-segment consumer heterogeneity. By placing mass on different components, the Dirichlet process considers across-segment heterogeneity on the one hand, while accounting through a variety of b_{l_n} and W_{l_n} for thick tails and skewed distributions on the other. $\alpha_{DPP} \in \mathbb{R}$ is referred to as the *concentration parameter* or *Dirichlet process tightness parameter*, and affects both the amount of across-segment consumer heterogeneity, and the approximation of heavy-tailed or skewed distributions by influencing the number of components. G_0 can be interpreted as a base measure (described below). By increasing α_{DPP} we can place a higher prior probability on models with a large number of components (Rossi 2014, pp. 72–76). We chose a flexible prior for the concentration parameter based on Conley et al. (2008):

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

The advantage of this prior (as compared to e.g. gamma priors) is that implications for the distribution of the number of possible components are more intuitive to assess (for more details see e.g. Rossi 2014, pp. 72–76). Hence, $\underline{\alpha} \in \mathbb{R}$ and $\overline{\alpha} \in \mathbb{R}$ were chosen to reflect the range of the probable number of components. ρ is a power parameter, which spreads out the prior mass. Escobar and West (1995) for example used a prior gamma distribution for α_{DPP}, while Ohlssen et al. (2007) and Voleti et al. (2017) applied a uniform distribution. Importantly, Voleti et al. (2017) examined the effect of different functional forms for the prior on α_{DPP}, finding that the corresponding 95 % credible intervals differed only marginally. Following Conley et al. (2008), the base distribution G_0 is parametrized as follows:

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

Within the Dirichlet process, the base distribution G_0 can be seen as a mean distribution, whereas α_{DPP} is a kind of variance of (b_{l_n}, W_{l_n}). Sethuraman (1994) provides a specification of the Dirichlet process prior in terms of the so-called stick-breaking representation. Here, the draws from the Dirichlet process can be represented as an infinite mixture of discrete vectors with specific probabilities following a beta distribution depending on α_{DPP}. The priors on a, ν and u are: $a \sim \mathrm{U}(a_l, a^u)$, $u \sim \mathrm{U}(u_l, u^u)$, $\nu \sim d - 1 + \exp(\zeta)$, $\zeta \sim \mathrm{U}(\nu_l, \nu^u)$, where d is the dimension of the data (number of mean part-worth utilities), and U is the uniform distribution.

2.3. Model estimation

We applied Markov Chain Monte Carlo (MCMC) methods to take draws from the posterior distributions of the HB-MNL, the MoN-HB-MNL, and the DPM-HB-MNL models. The draws for the individual part-worth utilities were generated by an improved Metropolis-Hastings random walk method with increments, whose covariance matrix can be tuned via a scaling parameter to approximate the conditional posterior best possible (following Rossi 2014, pp. 159–161 and Rossi et al. 2005, pp. 133–136). This covariance matrix among other things depends on the so-called fraction likelihood of respondent n which is used to compute the Hessian, and results from a multiplicative function of the MNL unit likelihood and the pooled likelihood. In this approach (referred to as method (iii) in Rossi et al. 2005, p. 137), a fractional likelihood parameter determines the weights of both likelihoods (Rossi 2014, p. 160). The fractional likelihood approach produces a higher degree of consumer heterogeneity (compared to standard HB models as seen e.g. in Train and Sonnier 2005) [2], and is implemented by default in the R code of the bayesm package, which we used for estimation of all HB models considered here. For more details, compare Rossi et al. (2005) and Rossi (2014).

Based on Rossi (2014, pp. 16–25), we chose the following diffuse prior configuration for our application: $k = d + 3$, $w = .01$, $\overline{b} = 0$, $\alpha = (5,...,5)^T$, $\Sigma = \mathrm{kI}$, $\overline{\delta} = 0$, $A_\delta = .01\mathrm{I}$, where d is the dimension of the data (here, the number of mean part-worth utilities), and I is the identity matrix. For further analysis, we set the fractional likelihood parameter to 1 (i.e. only the pooled likelihood was used to compute the Hessian) [2]. We tested different settings regarding the fractional likelihood parameter, finding that it only had a marginal impact on our measures of performance when using the diffuse prior settings of Rossi (2014, pp. 16–25). Following Rossi (2014, p. 29) again, we started with a large number of components (here: $M = 9$) for estimation of the MoN-HB-MNL model, and allowed the sampler to shut down a number of the components in the posterior, which corresponds to an implicit model selection step (compare the empirical study in Section 3). As a result, only those components were added to the posterior that provide additional flexibility to approximate the density shape. Alternatively, estimation of the MoN-HB-MNL could be carried out for a fixed number of $M \in \{1,...,C\}$ components, followed by an explicit model selection procedure. For example, model selection can then be based on the 95 %-trimmed log marginal likelihood suggested by Dubé et al. (2014) [3]. The marginal likelihood penalizes models that have a higher complexity (i.e. a larger number of estimated parameters) more strongly (Rossi 2014, p. 168), and is a well-established measure for ("explicit") model selection to favour more parsimonious models. Note that the MoN-HB-MNL includes the HB-MNL as a special case when setting the number of components to $M = 1$.

The DPM-HB-MNL prior parameter settings were chosen following Conley et al. (2008) and Rossi (2014, pp. 81–89). We set $\rho = .8$ and the other prior parameters to $a_l = .01$, $a^u = 10$, $u_l = .1$, $u^u = 4$, $\nu_l = .01$, and $\nu^u = 3$. $\underline{\alpha}$ and $\overline{\alpha}$ were chosen to provide a broad prior support for values from 1 to 50 components. Since the choice of a prior on α_{DPP} is of particular interest (as the expected number of components depends on α_{DPP} (Antoniak 1974)), we conducted a sensitivity analysis for these prior settings of the DPM-HB-MNL model. The results (part-worth utilities, goodness-of-fit measures, and the number of components) differed only marginally for different choices of $\underline{\alpha}$ and $\overline{\alpha}$, which is in line with the findings reported in Voleti et al. (2017).

The MCMC sampler was run for 210,000 iterations with a burn-in period of 110,000 iterations. To reduce possible correlation of the draws and prevent internal storage problems, we only used every 100th draw of the remaining 100,000 draws. We then used each of the 1,000 saved draws of the posterior distributions after the burn-in period to account for uncertainty in individual part-worth utilities. More precisely, each performance measure (see Subsection 2.4) was computed based on the draw level. We subsequently computed 95 % credible intervals of the resulting distributions. In a recent publication, Hein et al. (2022) could show for both simulated and empirical data that using the individual draws from an estimated HB model considerably improves the accuracy of shares of choice predictions in market simulations compared to using point estimates or other options. In addition, when applying a Bayesian approach for parameter estimation, it is theoretically correct to compute related quantities (such as choice shares or other performance measures) in a fully Bayesian manner, i.e. based on individual draws rather than on point estimates obtained from previously averaging individual draws (Hein et al. 2022). To ensure the convergence of the Markov chains, we monitored time-series plots and checked whether goodness-of-fit measures oscillated only randomly around their final values. Each check demonstrated that all MCMC chains reached stable states.

2.4. Measures of performance

Estimation of the HB-MNL, MoN-HB-MNL, and DPM-HB-MNL models based on CBC data provides us with individual part-worth utilities. We assessed the statistical performance of the three different types of models based on several performance measures, using the percent certainty (PC), the root likelihood (RLH), and the in-sample hit rate (IHR) as goodness-of-fit measures, as well as the out-of-sample hit rate (OHR) to evaluate the predictive model performance. Because no holdouts were collected for our data at hand, we performed leave-one-out cross-validation. A more detailed description is provided below.

The percent certainty, also known as likelihood-ratio index, pseudo R^2, or McFadden's R^2, compares the (final) log likelihood of a model to the likelihood of the null model, i.e. a model where all covariate effects are assumed to be zero ($\beta\; = (0,...,0)^T$) (Hauser 1978; McFadden 1977; Ogawa 1987):

$$PC(\beta^r) = \frac{LL^r_{final} - LL_{null}}{-LL_{null}}, \tag{7}$$

where LL^r_{final} and LL_{null} denote the (final) log-likelihood of the considered model based on draw r and the null log-likelihood, respectively. Values between even .2 and .4 indicate a satisfactory fit (McFadden 1977).

The log-likelihood is calculated by:

$$\mathrm{LL}^{\mathrm{r}} = \ln(\mathrm{L}(\beta^r)) = \sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{s=1}}^{S_n} \sum_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{Y}_{\mathrm{njs}} \ln(P^r_{njs}), \tag{8}$$

where S_n describes the choice sets offered to respondent n. $\mathrm{Y}_{\mathrm{njs}}$ is a dummy variable indicating whether respondent n has selected alternative j from choice set s (= 1) or not (= 0), and P^r_{njs} is the choice probability of respondent n for alternative j in choice set s based on the r-th draw.

The root likelihood represents the geometric mean of hit probabilities (e.g. Jervis et al. 2012; Sawtooth Software 2017):

$$\mathrm{RLH}(\beta^r) = \sqrt[\mathrm{NS_n}]{\prod_{\mathrm{n=1}}^{\mathrm{N}} \prod_{\mathrm{s=1}}^{\mathrm{S_n}} \prod_{\mathrm{j=1}}^{\mathrm{J}} {P^r_{njs}}^{\mathrm{Y_{njs}}}}. \tag{9}$$

As a rule, the value range of the RLH is between the reciprocal of the number of alternatives in a choice set (here: 1/J) and 1. The lower bound is reached for equal choice probabilities of all alternatives, which is synonymous with equal deterministic utilities for all alternatives, and hence corresponds to the situation in the null model (if all effects are zero, the utilities of all alternatives are the same). In other words, the RLH of the null model serves as a benchmark against which the RLH of the estimated model should be assessed.

The in-sample hit rate reflects the share of first-choice hits in the estimation sample (e.g. Andrews et al. 2002b; Voleti et al. 2017). A hit occurs if a respondent has actually chosen the alternative with the highest deterministic utility from a choice set as computed from the estimated model. This requires a respondent to have truly chosen the alternative with the highest utility from the choice set, which is consistent with random utility theory and all models considered here. Again, the chance hit rate (here: 1/J) serves as benchmark for assessing the model fit.

We further used the out-of-sample hit rate to evaluate the predictive performance of the models. Since no additional holdout choice tasks were collected for the data at hand (as mentioned above), we randomly selected one of the choice tasks evaluated by each respondent as a holdout to generate a validation sample. This means we estimated each model for $\mathrm{S_n} - 1$ choice tasks per respondent, leaving out each one randomly determined choice task, and applied the estimated model to predict the choices of respondents for the holdouts. Again, the reciprocal value of the number of alternatives in the holdout choice task serves as a benchmark for the realized out-of-sample hit rate.

3. Case study

The following applies the different types of models (HB-MNL, MoN-HB-MNL, DPM-HB-MNL) to an empirical CBC data set. First, we describe the relevant attributes and attribute levels used in our empirical study. We then report and compare the statistical performance of the three models with regard to goodness-of-fit and predictive accuracy, and show how the estimated heterogeneous preference structures can be interpreted. We also focus on cross-country heterogeneity, specifically analysing the benefit of including observable respondent characteristics (country dummies) as concomitant variables in the models.

3.1. Data

The data for our empirical study was provided by Kantar (TNS), one of the largest market research institutes in Germany and worldwide. The data was drawn from the product category *summer tires* and comprised 4,026 respondents. It was collected in 2016 in France (*n* = 820), Germany (*n* = 802), Spain (*n* = 800), Italy (*n* = 804), and the United Kingdom (*n* = 800). The focal product of summer tires was described by 17 relevant attributes, among them a brand attribute representing 10 different competitors. A special feature of the data was that brand-specific price attributes were used, i.e. each brand was characterized by an own unique set of price levels. The use of alternative-specific price levels allowed different price quality tiers of summer tires to be adequately included. Accordingly, 10 out of the 17 attributes were price attributes, and an alternative-specific design was used to build the choice tasks for the respondents. The attributes and attribute levels used in the study are shown in *Tab. 1*.

The latter levels of the two attributes *rolling resistance/fuel consumption* and *grip on wet roads* refer to the EU tire labels required by the European Union, and indicate the performance (efficiency) of a tire regarding them [4]. The star levels of the two attributes *consumer reviews* and *independent test results* correspond to the established 5-star rating system widely used for product or service evaluations. Here, *not available* was included as additional level to simulate the situation in real world settings where a rating may not always be available.

Each respondent cycled through 15 choice sets and was asked each time to choose the most preferred summer tire, resulting in a total of 60,390 observations. The choice sets consisted of three alternatives plus a "no purchase" option. The first three positions in the choice sets representing the "real" alternatives received almost the same choice shares (26.54 % for position 1, 26.24 % for position 2, and 26.09 % for position 3), suggesting a very well-balanced, randomly generated choice task design. The "none" alternative obtained a choice share of 21.13 %, which is a slightly higher share compared to what is commonly observed in empirical CBC studies with many attributes and attribute levels. According to Johnson and Orme (1996), typical shares for the "none" option lie between 5 % and 15 %.

We estimated the HB-MNL model, the MoN-HB-MNL model (with nine components), and the DPM-HB-MNL model based on the 60,390 observations. Because the data did not contain fixed holdout tasks, we randomly split the choice sets into an estimation sample (14 choice sets per respondent; 56,364 observations) and a validation sample (1 choice set per respondent; 4,026 observations). Since respondents may need some time for orientation or adaption in the initial phase of the choice task and/or may fatigue in later choice tasks, the holdout was selected at random for each respondent out of the middle 50 % of the 15 choice sets, i.e. out of choice sets 5 to 11. We repeated this validation procedure a second time to reduce variability and to compensate for unwanted systematic effects that could have been caused by a one-time random split. In this context, we could have further performed leave-one (choice set)-out cross validation to minimize potential selection biases, which would have required conducting 15 rounds of cross-validation using each of the 15 choice sets per respondent once for validation. However, we limited ourselves to two rounds of cross-validation due to the high number of observations and the resulting long computation times. We averaged the out-of-sample hit rates (based on draw-level) over the two rounds of cross-validation to assess the predictive performance of the models. We used dummy-coding (except for the price attributes) for model estimation by specifying all first attribute levels as reference categories (i.e. "Michelin", "Basic", "-8,000 km", "A", "A", "not available", "not available"). The brand-specific price attributes were coded linearly to stay parsimonious (to save degrees of freedom), leading to the estimation of a total of 44 part-worth utilities/parameters on an individual respondent level (including the "no purchase" parameter).

3.2. Results

We evaluated the predictive accuracy of the models by averaging the results of both validation samples, as mentioned above. All goodness-of-fit measures were computed based on the estimation results for the entire data set. As outlined in Section 2.3, we estimated the MoN-HB-MNL model with M = 9 components, allowing the MCMC sampler to shut down a number of the components rather than starting with a small number of components and applying an "explicit" model selection procedure. All results for the different performance measures (including 95 % credible intervals of the posterior distributions) obtained for the three models either with or without concomitant variables (referred to as z-variables) are summarized in *Tab. 2*.

Attribute	Attribute levels	# of attribute levels
Brand	• Michelin • Continental • Goodyear • Bridgestone • Pirelli • Kleber • Client Brand • Firestone (Germany: Uniroyal) • Hankook • Low price brand	10
Tire Type	• Basic • Comfort • Sports	3
Longevity	• -8,000 km • -4,000 km • Basic • +4,000 km • +8,000 km	5
Rolling resistance/ Fuel consumption	• A • B • C • E • F	5
Grip on wet roads	• A • B • C • E • F	5
Consumer reviews	• not available • 1 star • 2 stars • 3 stars • 4 stars • 5 stars	6
Independent test results	• not available • 1 star • 2 stars • 3 stars • 4 stars • 5 stars	6
Price Michelin	• 77 €, 81 €, 85 €, 89 €, 93 € , 97 €, 101 €	7
Price Continental	• 69 €, 73 €, 77 €, 81 €, 85 €, 89 €, 93 €	7
Price Goodyear	• 65 €, 69 €, 73 €, 77 €, 81 €, 85 €, 89 €	7
Price Bridgestone	• 65 €, 69 €, 73 €, 77 €, 81 €, 85 €, 89 €	7
Price Pirelli	• 65 €, 69 €, 73 €, 77 €, 81 €, 85 €, 89 €	7
Price Kleber	• 57 €, 61 €, 65 €, 69 €, 73 €, 77 €, 81 €	7
Price Client Brand	• 61 €, 65 €, 69 €, 73 €, 77 €, 81 €, 85 €	7
Price Firestone	• 53 €, 57 €, 61 €, 65 €, 69 €, 73 €, 77 €	7
Price Hankook	• 53 €, 57 €, 61 €, 65 €, 69 €, 73 €, 77 €	7
Price low-price brand	• 49 €, 53 €, 57 €, 61 €, 65 €, 69 €, 73 €	7

Tab. 1: Attribute and attribute levels used in the empirical study

Model	# of mixture components	LL	PC	RLH	IHR	OHR
HB-MNL	$M = 1$	[-28289.26;-27330.77]	[.662;.674]	[.626;.636]	[.809;.816]	[.515;.532]
MoN-HB-MNL	$M = 9^{a}$	[-28351.41;-27335.51]	[.661;.673]	[.625;.636]	[.808;.816]	[.515;.533]
DPM-HB-MNL	$[M = 1]^{b}$	[-38537.94;-37550.30]	[.540;.551]	[.528;.537]	[.736;.744]	[.533;.551]
HB-MNL (z-var)	$M = 1$	[-27831.58;-26857.83]	[.668;.679]	[.631;.641]	[.811;.819]	[.516;.533]
MoN-HB-MNL (z-var)	$M = 9^{a}$	[-27839.80;-26949.26]	[.667;.678]	[.631;.640]	[.811;.819]	[.515;.534]
DPM-HB-MNL (z-var)	$[M = 1]^{b}$	[-37970.41;-37095.06]	[.546;.557]	[.533;.541]	[.740;.747]	[.535;.551]

Notes: a: MoN-HB-MNL model estimated initially with nine components, allowing the components to be shut down in the posterior, and resulting in a one-component solution.
b: The number of components were obtained as a result a posteriori. The DPM-HB-MNL model returned one component for our data set (as indicated by [M = 1]) as well.

Tab. 2: Goodness-of-fit and predictive accuracy statistics by model type. Shown are the 95 % credible intervals of the posterior distributions.

We first discuss the results for the models that did not include the additional z-variables (i.e. that did not account for observed respondent heterogeneity). When analysing the posterior draws of the mixture probabilities of the MoN-HB-MNL model, we could observe that the mixture model with initially nine components degenerated into a one-component solution (i.e. to a model with a single normal component, as provided by HB-MNL by definition). Since credible intervals for all goodness-of-fit and predictive accuracy statistics overlap between the estimated MoN-HB-MNL solution and the HB-MNL solution, both model types provide comparable results for the data at hand; the additional flexibility of the MoN-HB-MNL does not pay off here.

The DPM-HB-MNL model results in a one-component solution as well. Because of the different (more flexible) prior settings of the DPM-HB-MNL model, goodness-of-fit and predictive accuracy results turn out different nevertheless: goodness-of-fit statistics (LL, PC, RLH, and IHR) are much worse than for the HB-MNL and MoN-HB-MNL solutions, while predictive validity measured in terms of OHR is superior (even if the upper bound value of the credible interval obtained for the MoN-HB-MNL model (.533) coincides with the lower bound value of the credible interval for the DPM-HB-MNL model (.533)). The differences in fit and predictive accuracy between models can be explained by a larger shrinkage effect of the DPM-HB-MNL model, as exemplarily displayed for the brand intercepts in *Fig. 1* (except for Michelin, which constitutes the reference category). The figure presents posterior means of the marginal densities for the brand intercepts, contrasting the HB-MNL model (solid line), the MoN-HB-MNL model (dotted line), and the DPM-HB-MNL model (bold solid line). While the differences in fitted densities between the HB-MNL model and the MoN-HB-MNL model are rather small, the distributions obtained for the DPM-HB-MNL are clearly steeper and more centered around the population mean. There, individual part-worth utilities are shrunk more strongly towards the population means compared to the other two models [5]. On the other hand, we found high Pearson correlations between estimated individual part-worth utilities of the HB-MNL and MoN-HB-MNL models (.97), HB-MNL and DPM-HB-MNL models (.97), and MoN-HB-MNL and DPM-HB-MNL models (.95), indicating that differences between the choice models do not appear substantial [6] (compare *Tab. 3*, left upper section).

We augmented the three models to consider the potential impact of the country of origin of the respondents on preference structures and model performance. We specifically included country dummies for the respondents in the upper level of the models as additional predictors (z-variables) for part-worth estimation to capture observed cross-country heterogeneity (compare equation 3).

All performance measures displayed in *Tab. 2* suggest only very little benefit from this extension. Goodness-of-fit and predictive accuracy measures between choice models with and without country dummies are comparable, as indicated by the overlapping credible intervals. Note that credible intervals for the predictive model performance measured by OHR not only overlap, but are almost identical. On the other hand, a closer look at the Δ matrix of the extended HB-MNL model (i.e. including z-variables, compare equation 3) shows that in 70 of the 176 cases (remember that Δ is a (n_z,d)-matrix, with n_z and d denoting the number of country dummies (= 4) and the number of part-worth utilities (= 44)), the corresponding 95 % credible intervals do not include the 0, indicating a relationship between country of origin and the respondents' (mean) part-worth utilities. The same applies to the extended MoN-HB-MNL and DPM-HB-MNL models for 69 or 72 of the 176 cases. Nevertheless, the general impression is one of a weak relationship between the country of origin and attribute preferences, which is not only visible from the performance measures (*Tab. 2*) but also from the Pearson correlations for individual part-worth utilities between models with versus without country dummies, which are consistently above .94 (*Tab. 3*). Similar to *Fig. 1*, we also compared the posterior means of the marginal densities between the models with and without z-variables. The distributions here

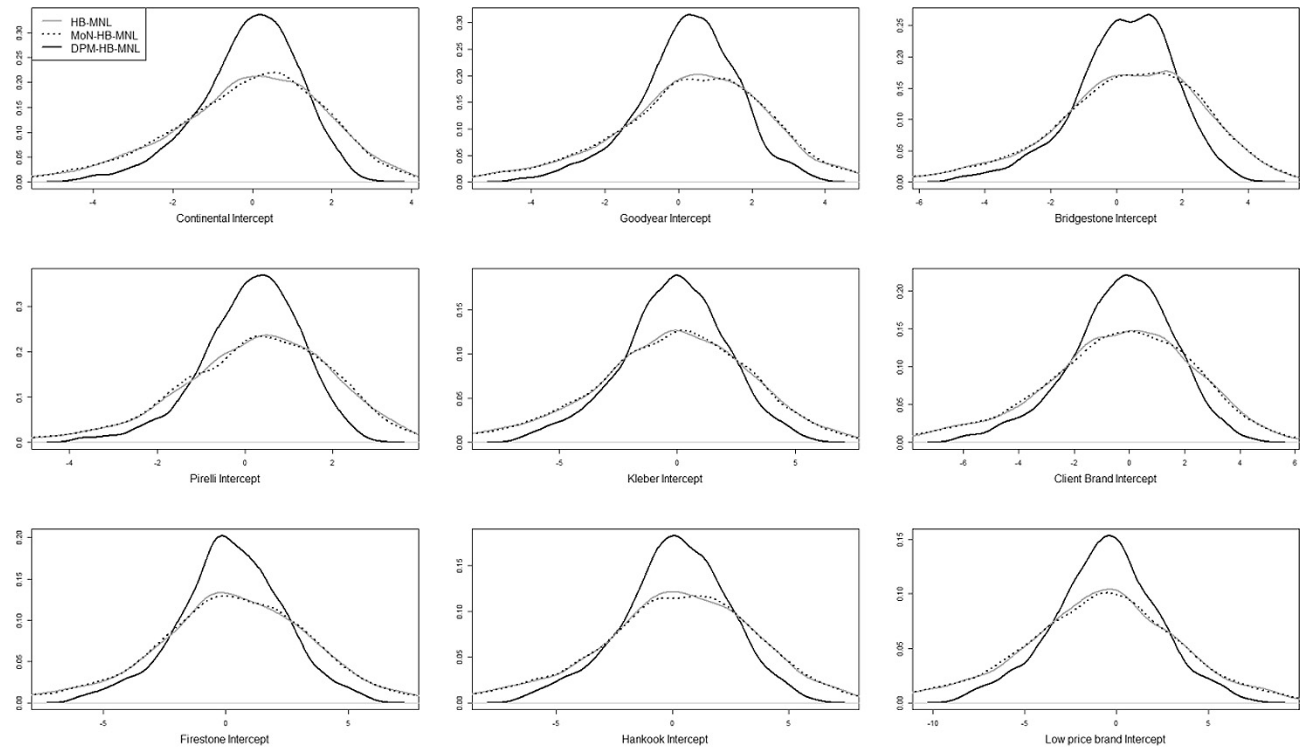


Fig. 1: Posterior means of marginal densities for brand intercepts

	HB-MNL	MoN-HB-MNL	DPM-HB-MNL	HB-MNL (z-var)	MoN-HB-MNL (z-var)	DPM-HB-MNL (z-var)
HB-MNL	1	.97	.97	.99	.97	.96
MoN-HB MNL		1	.95	.96	.94	.94
DPM-HB-MNL			1	.96	.94	.98
HB-MNL (z-var)				1	.97	.97
MoN-HB-MNL (z-var)					1	.95
DPM-HB-MNL (z-var)						1

Tab. 3: Pearson correlations between individual part-worth utilities

are nearly identical and are available from the authors upon request.

Overall, including country of origin covariates for respondents does not improve the predictive model performance, nor does it make a difference in terms of pairwise correlations between individual part-worth utilities or posterior means of marginal densities, as discussed above. The fact that both the MoN-HB-MNL model and the DPM-HB-MNL model (as the HB-MNL model by definition) also only suggest a 1-component solution further underscores how cross-country heterogeneity does not play an important role for the data at hand.

Which of the three models should be selected based on the results so far? Following van Heerde et al. (2002), marketing managers should rely on the model with the best predictive performance, which here is the DPM-HB-MNL model (both with and without z-variables). On the other hand, differences in OHR between the DPM-HB-MNL and the two other types of models (HB-MNL, MoN-HB-MNL) are not that large (only about 2 %) for our data. Furthermore, the HB-MNL model is the most parsimonious model and is currently the only model that is implemented in commercial software packages (e.g. Sawtooth Software), the latter of which facilitates its application for practitioners. Finally, extending our models by concomitant variables (here the country of origin dummies) to account for observed heterogeneity between respondents did not pay off, at least not from a statistical perspective (compare *Tab. 2*) [7]. Nevertheless, using the country of origin information for the respondents in a post-hoc segmentation allowed us to identify a few differences in brand preferences between respondents from different countries (see the discussion in Subsection 3.3 below). For the reasons mentioned, we will continue comparing and interpreting the part-worth utility structures obtained for the HB-MNL model (most parsimonious model) and the DPM-HB-MNL model (best predictive performance) in the next section, and abstain from inspecting more closely the results for the MoN-HB-MNL model as well as for the models with concomitant variables. In the online appendix, we summarize and briefly discuss goodness-of-fit and predictive validity statistics for the three choice models (HB-MNL, MoN-HB-MNL and DPM-HB-MNL models without z-variables) when estimated with part-worth utility functions as well for the brand-specific price attributes (leading to the estimation of a total of 94 part-worth parameters on an individual respondent level compared to 44 part-worth parameters in case the price attributes were coded linearly, see Section 3.1).

3.3. Interpretation of heterogeneous preference structures

As discussed, the DPM-HB-MNL model provided a higher cross-validated hit rate compared to the HB-MNL model, while the latter showed better goodness-of-fit statistics caused by a weaker shrinkage effect. *Tab. 4* displays estimated posterior means and standard deviations for both models on the individual attribute level. First, we observe a very similar part-worth utility structure estimated by the two models. An average consumer (if one actually existed) would prefer a Goodyear or Bridgestone tire in its basic version, with the highest longevity (+8,000 km), optimal rolling resistance/fuel consumption (EU tire label "A"), optimal grip on wet roads (EU tire label "A"), and with the best ratings in consumer reviews and independent tests ("5 stars"). We further observe that posterior means of price effects are negative for all brands except for the low-price brand, which was the cheapest brand in the tire market at the time the data was collected. For the low-price brand, the posterior mean price effect approaches zero in both models (.01 in the HB-MNL and .00 in the DPM-HB-MNL). As expected, a (nearly) monotonic increase (the more the better) or decrease (the less the better) of mean part-worth utilities for attribute levels is obtained for: *longevity*, *rolling resistance/fuel consumption*, *grip on wet roads*, *consumer reviews*, and *independent test results*. Here, the best attribute levels are the most preferred ones across respondents (as indicated by the population mean part-worths in *Tab. 4*). Not unexpectedly, very bad consumer reviews or test results ("1 star") are evaluated as being even worse than if no consumer reviews or test results were available at all.

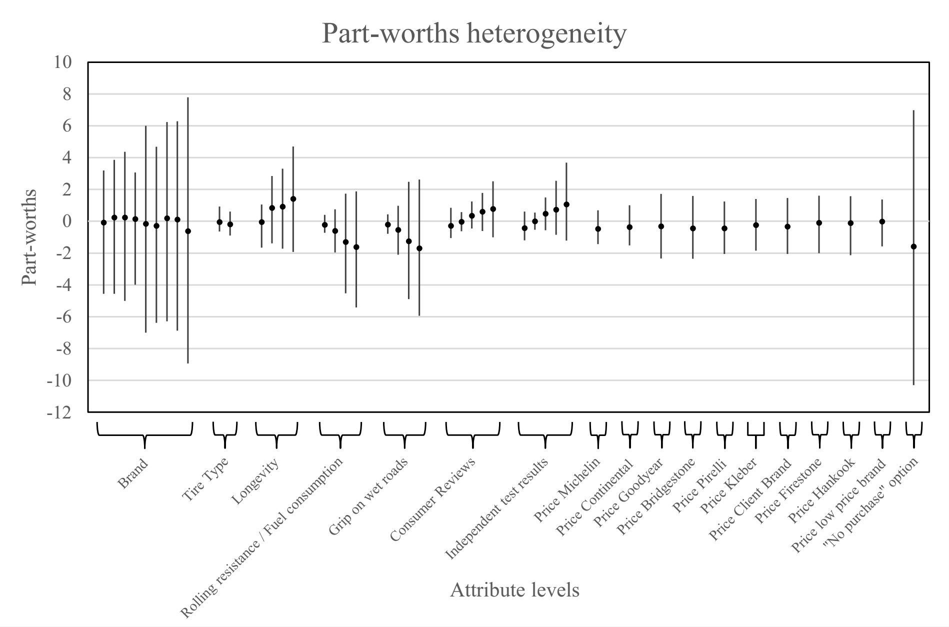
Fig. 1 showed a stronger shrinkage effect of the DPM-HB-MNL model compared to the HB-MNL model for all estimated brand intercepts. In *Tab. 4*, this stronger shrinkage of the DPM-HB-MNL model towards the population means is reflected by consistently smaller standard deviations across all attributes and levels. This explains the conflicting results of a better predictive performance, but a worse fit of the DPM-HB-MNL model compared to the HB-MNL model, suggesting overfitting of the HB-MNL model. In the following, we focus on interpreting the heterogeneous preference structures of the DPM-HB-MNL model in greater detail.

Based on the arrangement of attributes and attribute levels in *Tab. 4*, *Fig. 2* illustrates the estimated distributions of individual-level part-worths obtained for the DPM-HB-MNL model (not included are the reference levels for the categorical attributes) by depicting the spans of individual part-worth estimates for each of the attributes and levels. This figure shows a large amount of heterogeneity in utility structures across respondents, as well as large differences in the amounts of heterogeneity across attributes and their levels. The largest ranges of individual part-worth utilities (apart from the "no purchase" parameter) can be observed for the *brand* attribute representing the different competitors. It is noticeable that the brands out of the "top 5 strongest brands" list (Bridgestone, Pirelli, Continental, Goodyear) (Brand Finance 2020, p. 21) show clearly smaller amounts of heterogeneity (maximum part-worth: 4.37, minimum part-worth: -5.01) compared to the other brands (Kleber, Client Brand, Firestone, Hankook, and the low-price brand) with part-worths in the range between -8.93 and 7.80. For the attributes *rolling resistance/fuel consumption* and *grip on wet roads*, we obtained a larger amount of heterogeneity across respondents for the worst levels (EU tire labels "E" and "F"). In contrast, respondents' preferences differ more strongly for better levels of the attributes *longevity (+8,000 km)*, *consumer reviews (5 stars)*, and *independent test results (5 stars)*.

Fig. 3 displays different utility structures of three selected respondents (respondents 1303, 1532, and 3494) for two of the tire attributes together, with the preference structure for an average consumer resulting from the DPM-HB-MNL model. The latter is represented by the mean of the individual part-worth utilities of respondents and therefore ignores preference heterogeneity. The brand preference structures of respondents 1532 and 3494 are completely different. Whereas respondent 3494 clearly favoured the lower-priced non-premium brands (Kleber, client brand, Firestone, Hankook, and the low price brand), respondent 1532 apparently tended towards the higher-quality premium brands ("top 5 strongest brands"), with by far the highest utility attributed to Michelin. Both respondents also largely differ in their brand preferences from the (fictitious) average consumer, who appears to be relatively indifferent to the various brands (except for the low-price brand which is somewhat less preferred). This is clearly the result of averaging out the highly heterogeneous individual brand preference patterns. In contrast, respondent 1303 showed a much less clearer preference pattern with regard to brand utilities, even if she/he also happened to prefer the brands Kleber, Hankook, and the low-price brand least, similar to respondent 1532. On the other hand, respondent 1303 placed more weight on the technical feature of grip on wet roads, clearly rejecting the two worst tire labels E and F, while the utility patterns of the other two respondents (1532, 3494) were close to that of the average consumer. Nevertheless, all respondents preferred tires with a higher performance regarding the technical attribute, where utility consistently decreased with a less efficient tire label. Overall, the three respondents considerably

		HB-MNL		DPM-HB-MNL	
Attribute	Attribute levels	Posterior mean	Std Dev	Posterior mean	Std Dev
Brand	Michelin	.00	.00	.00	.00
	Continental	-.13	1.97	-.07	1.25
	Goodyear	.35	2.11	.24	1.37
	Bridgestone	.35	2.29	.24	1.52
	Pirelli	.24	1.77	.15	1.13
	Kleber	-.23	3.30	-.16	2.23
	Client Brand	-.41	2.76	-.28	1.88
	Firestone	.30	3.17	.20	2.14
	Hankook	.18	3.40	.11	2.31
	Low price brand	-.87	4.19	-.61	2.86
Tire Type	Basic	.00	.00	.00	.00
	Comfort	-.06	.43	-.04	.19
	Sports	-.27	.45	-.18	.20
Longevity	-8,000 km	.00	.00	.00	.00
	-4,000 km	-.09	.70	-.03	.35
	Basic	1.17	1.22	.84	.79
	+4,000 km	1.30	1.48	.93	.94
	+8,000 km	1.99	1.73	1.41	1.16
Rolling resistance / Fuel consumption	A	.00	.00	.00	.00
	B	-.31	.39	-.21	.18
	C	-.84	.78	-.59	.50
	E	-1.84	1.76	-1.30	1.21
	F	-2.28	2.11	-1.60	1.46
Grip on wet roads	A	.00	.00	.00	.00
	B	-.28	.41	-.19	.21
	C	-.76	.85	-.53	.55
	E	-1.80	1.92	-1.25	1.32
	F	-2.40	2.30	-1.69	1.61
Consumer reviews	not available	.00	.00	.00	.00
	1 star	-.37	.56	-.27	.34
	2 stars	-.04	.43	-.03	.22
	3 stars	.52	.47	.36	.25
	4 stars	.86	.71	.61	.42
	5 stars	1.12	.95	.79	.58
Independent test results	not available	.00	.00	.00	.00
	1 star	-.62	.55	-.42	.32
	2 stars	-.02	.39	.00	.16
	3 stars	.64	.60	.48	.35
	4 stars	1.04	.88	.74	.55
	5 stars	1.52	1.25	1.07	.82
Price Michelin		-.71	.60	-.47	.35
Price Continental		-.53	.69	-.36	.42
Price Goodyear		-.45	.98	-.32	.65
Price Bridgestone		-.62	.94	-.43	.61
Price Pirelli		-.61	.81	-.44	.52
Price Kleber		-.33	.82	-.23	.51
Price Client Brand		-.47	.86	-.33	.55
Price Firestone		-.13	.88	-.09	.59
Price Hankook		-.18	.86	-.11	.59
Price Low-price brand		.01	.75	.00	.49
No purchase		-1.67	4.50	-1.58	3.39

Tab. 4: Distributions of individual-level part-worths. Posterior means and standard deviations (HB-MNL vs. DPM-HB-MNL model)



Notes: The length of a bar corresponds to the range of the estimated part-worths for a respective attribute level and reflects the amount of heterogeneity across respondents. The estimated population mean part-worths are highlighted with a dot.

Fig. 2: Distributions of individual-level part-worths (DPM-HB-MNL)

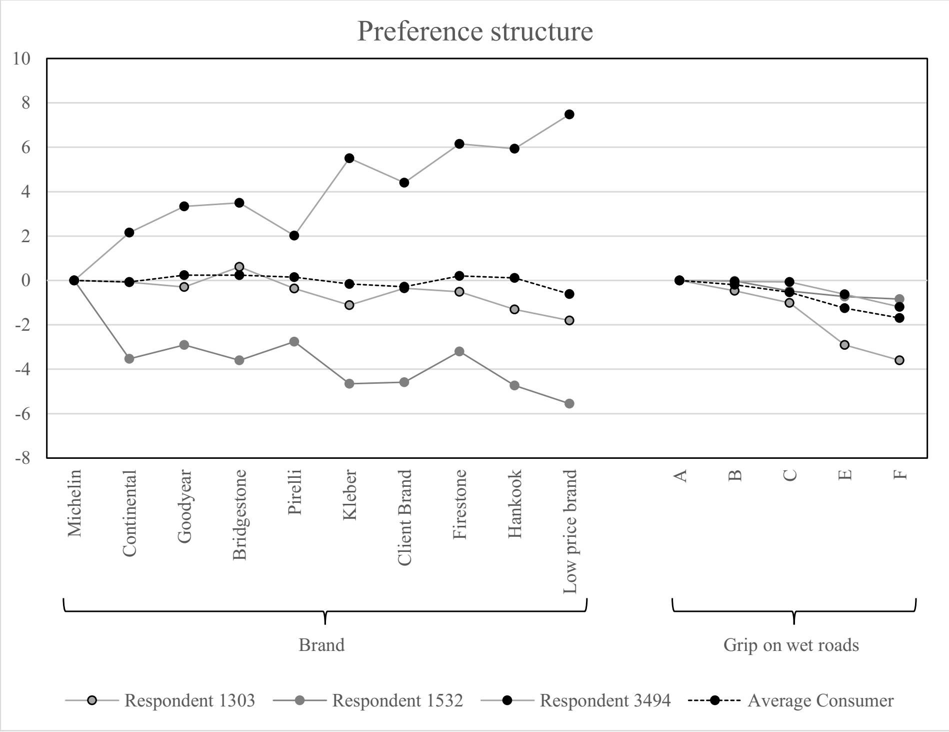


Fig. 3: Preference structure for the attributes brand and grip on wet roads of three selected respondents (DPM-HB-MNL)

differed in their preferences, especially in their part-worth utility patterns for the different brands (high preference for lower-priced non-premium versus higher-priced premium brands). Therefore, ignoring existing preference heterogeneity (i.e. assuming an average consumer) might obscure important aspects, and would most likely lead to biased predictions and/or wrong managerial implications.

In Section 3.2, we concluded that the inclusion of additional covariates (referred to as concomitant or z-variables) accommodating a possible country of origin effect on respondents' preference formation did not substantially pay off for our data, especially not with regard to the predictive model performance. Nevertheless, we could observe a slight increase in goodness-of-fit statistics related to a number of significant country of origin effects on mean part-worth utilities (which of course could also be at least partly driven by the large sample size). In other words, even if the model performance could not be noticeably improved, a small part of the heterogeneity in respondents' preferences can still be explained by or assigned to these covariates. For this reason, it appears worthwhile to take a final look at potential differences

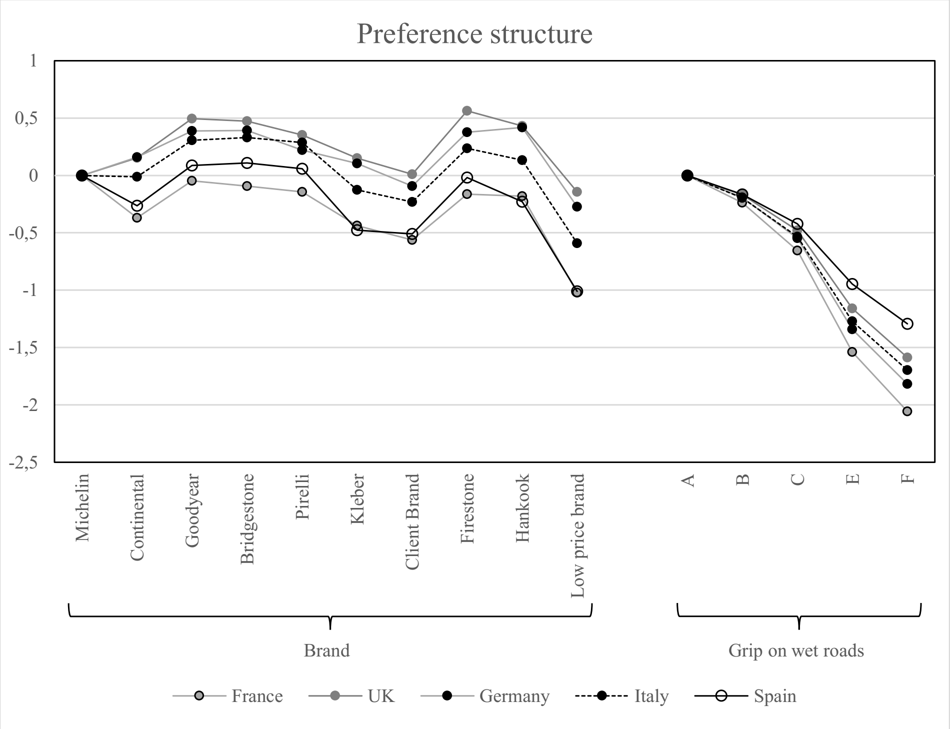


Fig. 4: Cross-country heterogeneity for the attributes brand and grip on wet roads (DPM-HB-MNL)

between countries resulting from a post-hoc segmentation we conducted. In an effort to remain consistent with the presentation of the results in this section, and because of the highly similar part-worth utility structures obtained from the models with versus without concomitant covariates, we continue to focus on the results of the DPM-HB-MNL model without z-variables. Interpretation of the part-worth utilities of the models with concomitant variables by country is nearly identical; the corresponding results are available from the authors upon request.

Fig. 4 displays the different mean utility structures of the five different countries (France, UK, Germany, Italy, and Spain) for the attributes *brand* and *grip on wet roads*. The brand preference structures turn out very similar for the five countries. The client brand and the low price brand are least preferred in general, whereas the brands Goodyear, Bridgestone, Pirelli, Firestone, and Hankook are consistently among the brands with the highest part-worth utilities across the five countries. But we can also observe an important difference for brand preferences between the countries. The most preferred brand in France (on average) is Michelin, which served as the reference category for estimation with a corresponding part-worth utility of 0. In contrast, the majority of the other brands are preferred to Michelin in the UK and Germany (except the low-price brand in the UK and Germany, and the client brand in Germany). Interestingly, brand preferences in Italy nearly coincide with the mean brand preferences across the five countries (not shown in *Fig. 4*, compare the posterior means for the DPM-HB-MNL model in *Tab. 4*). Respondents' preferences for the attribute *grip on wet roads* barely differ between the countries, and show the expected pattern: tires providing a better performance on wet roads are consistently preferred to tires with less efficiency. Similarly, consumer preferences for all other technical features (*tire type*, *longevity*, and *rolling resistance/fuel consumption*) as well as for the two attributes *consumer reviews* and *independent test results* differ only marginally between the five countries, and consequently nearly coincide with the pooled preference patterns, as given by the population means in *Tab. 4* [8].

In summary, we can observe a large amount of (unobserved) preference heterogeneity in our empirical data on the one hand, but only a very moderate amount of (observed) cross-country heterogeneity on the other. Perhaps as a consequence of missing substantial differences in preferences between the countries, both the MoN-HB-MNL model and the DPM-HB-MNL model suggest one-component solutions (as the HB-MNL model does by definition). The DPM-HB-MNL model revealed the highest degree of shrinkage and the best predictive performance across the three types of models (independent of whether the concomitant covariates were included or not). A second reason for the small amount of cross-country consumer heterogeneity may be attributed to the nature of the attributes used in the conjoint study. Except for the brand attribute and the type of tire, all other attributes are of a "the more the better" or "the less the better" nature, suggesting a natural (i.e. monotonic) order of preference (which we actually obtained from our models). Nevertheless, considering the potential impact of consumer background characteristics on part-worth utilities made it possible to identify at least some cross-country heterogeneity in terms of differences in brand preferences (as illustrated for the one-component DPM-HB-MNL model in *Fig. 4*). These differences can be explained for example by a county-of-origin effect, or by an effect of ethnocentrism and patriotism on consumer preferences, as captured by the country dummies (e.g. for Michelin).

4. Conclusion

Considering heterogeneity in consumers' choice behaviour has become a key focus of CBC studies. For more than two decades, ongoing research has examined how to model heterogeneity in conjoint approaches. One main focus here lies on the comparison of the performance of established competing approaches to capture consumer heterogeneity (e.g. Krueger et al. 2018; Otter et al. 2004; Voleti et al. 2017). Very popular among researchers and managers is the standard HB-MNL model framework (Allenby and Ginter 1995; Lenk et al. 1996), which is characterized by a unimodal continuous distribution to model consumer heterogeneity (referred to as HB-MNL in this article). Commercial software for the HB-MNL model is available, which is the primary reason for its prominence in marketing research and practice.

Researchers have recently applied and discussed HB choice models that are more flexible and capable of representing multimodal continuous heterogeneity (Baumgartner and Steiner 2007; Chen et al. 2017; Krueger et al. 2018; Löffler and Baier 2015; Voleti et al. 2017). MoN-HB-MNL and DPM-HB-MNL models allow for this kind of flexible representation of consumer heterogeneity and can also accommodate heavy-tailed and skewed distributions. The advantage of using a DPM-HB-MNL model is that the number of components does not need to be specified a priori. The Dirichlet process prior of the DPM model directly determines the number of mixture components based on the data and prior settings.

This contribution provided an overview of the most relevant and recent choice models for addressing continuous heterogeneity in consumers' preferences. We applied these models to an empirical data set, assessed their comparative performance in terms of goodness-of-fit and predictive accuracy, and interpreted the estimated heterogeneous preference structures. We further showed how to include consumer background characteristics in the upper level of these models as concomitant covariates to account for observed consumer heterogeneity, with a special focus on cross-county heterogeneity in our empirical application.

For our data, we observed that all choice models (with and without concomitant variables) resulted in a one-component solution. Due to more flexible prior assumptions, the DPM-HB-MNL model yielded a higher cross-validated hit rate compared to the MoN-HB-MNL and the HB-MNL model. The two latter models tended to slightly overfit the data, which was indicated by higher goodness-of-fit statistics and a lower predictive accuracy. We showed that this result could be attributed to the weaker extent of Bayesian shrinkage of these two models. Accordingly, the DPM-HB-MNL model is not only able to account for heavy-tailed distributions (as is known when using a sufficient number of normal components) but also for distributions with thinner tails, the latter of which was shown in our empirical study as a result of a larger shrinkage effect compared to the HB-MNL model (compare *Fig. 1*). Including concomitant covariates in terms of country of origin information for the respondents did not improve the statistical model performance (especially not the predictive performance), although it helped to explain a few differences in brand preferences between the five countries (this also applied where we did not use these consumer covariates in the model estimation step, but for post hoc segmentation by country instead).

The estimated posterior distributions further revealed a large amount of (unobserved) heterogeneity between respondents, especially with regard to (1) their brand preferences (as reflected by the large standard deviations for the levels of the *brand* attribute, see *Tab. 4*), but also for (2) worse levels of the technical features *rolling resistance/fuel consumption*, *grip on wet roads*, and better levels of the attribute *longevity* (also see the corresponding standard deviations in *Tab. 4*). This underpins how ignoring preference heterogeneity between consumers can harbour the danger of modelling an "average consumer" who simply does not exist in real markets, very likely leading to biased predictions and wrong managerial implications (e.g. for pricing and product design decisions, or for product positioning objectives).

Our results are only partially in line with Voleti et al. (2017) who compared various choice models (among them the HB-MNL, MoN-HB-MNL, and DPM-HB-MNL models) using eleven empirical data sets characterized by different numbers of observations, respondents, tasks, alternatives per task, attributes, and attribute levels. In their study, DPM-HB-MNL models outperformed the other models in terms of predictive accuracy as well. In addition, HB-MNL models did not perform worse than MoN-HB-MNL models in most of the data sets (as in our study). In one data set, the HB-MNL model even led to a better out-of-sample hit probability than the DPM-HB-MNL model. In contrast to our findings, the DPM-HB-MNL model also showed the best fit in their study. However, our empirical data set represents a much more complex scenario, i.e. it comprised a much larger number of attributes, respondents, observations, and parameters compared to all of the data sets used by Voleti et al. (2017). In addition, we demonstrated the performance of the models in an alternative-specific design context. The out-of-sample hit rates in our study turned out similarly high to those in Voleti et al. (2017).

Beyond fit and predictive validity, parameter recovery is also a relevant topic when investigating or comparing different choice models (e.g. Hein et al. 2019). Parameter recovery measures the fit between "true" and re-estimated parameters (here: part-worth utilities). However, since we used a real-life data set to illustrate the application of the different types of models, true preference structures of respondents were not known, and parameter recovery could not be assessed here. "True" preference structures are known in Monte Carlo simulations, allowing one to compare artificially generated "true" part-worth utilities with estimated part-worth utilities in this case. As Voleti et al. (2017) state, it is interesting to study the performance of all models under a reasonable distribution of heterogeneity. We expect that the results will depend on the assumptions about the heterogeneity distribution. Future research should address the comparison between these choice models under varying experimental conditions.

A final note regarding the CBC data set used for our empirical study: the data was comprised of respondents from five different countries and a very large number of attributes (17), and an alternative-specific design was chosen for data collection to allow for brand-specific price effects (each one linear price parameter for each of the 10 brands). Although from this perspective the data set can be classified as complex, both the DPM-HB-MNL and the MoN-HB-MNL resulted in a one-component solution (which we did not expect from scratch), as provided by the HB-MNL by definition. We mentioned several reasons for this at the end of Section 3.3 (very low cross-country heterogeneity, the majority of attributes were of a "the more the better" or "the more the worse" nature), which should have counteracted a larger diversity of preferences across respondents, and that did not allow the DPM-HB-MNL and MoN-HB-MNL models to develop their greater flexibility for uncovering more complex (e.g. multimodal) preference structures. On the other hand, even under these conditions, and the resulting minor differences between the part-worth utility estimates between models, the DPM-HB-MNL provided a somewhat better predictive performance. It furthermore seems from our results that the DPM-HB-MNL is somewhat more robust against overfitting, since the MoN-HB-MNL and the HB-MNL models provided a better model fit, albeit a worse predictive validity at the same time. More research is needed to validate this finding.

Notes

[1] An extensive discussion regarding the ongoing debate in the marketing literature on whether to address preference heterogeneity with a discrete or a continuous modelling approach can be found in Paetz and Steiner (2017) and Paetz et al. (2019). For the discrete approach, the use of latent class models to determine segments with homogeneous consumer preferences has become extremely popular (see in particular DeSarbo et al. 1995, Ramaswamy and Cohen 2007, and Wedel and Kamakura 2000). Teichert (2001) dealt very early in the German-language literature with the comparison of latent class and hierarchical Bayesian methods for utility estimation in choice-based conjoint analysis.

[2] Our thanks to an anonymous reviewer who pointed this out.

[3] There is a built-in function in bayesm for the log marginal likelihood (logMargDenNR). We thank an anonymous reviewer for this note.

[4] https://en.wikipedia.org/wiki/Tyre_label [05.10.2021]

[5] Density plots for the other attribute and attribute levels look very similar. A larger shrinkage effect can be observed for the DPM model in particular. All density plots are available from the authors upon request.

[6] Note that PC, RLH, IHR, and OHR statistics indicate a considerably better model performance compared to the null model for all three models.

[7] Note that including the four country dummies in our model requires the estimation of 44 x 4 = 176 additional effects captured by the Δ matrix.

[8] The country-level posterior mean part-worth utility structures for all other attributes not displayed in *Fig. 4* are available from the authors upon request.

References

Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating Prior Knowledge into the Analysis of Conjoint Studies. *Journal of Marketing Research*, 32(2), 152–162.

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

Allenby, G. M., & Ginter, J. L. (1995). Using Extremes to Design Products and Segment Markets. *Journal of Marketing Research*, 32(4), 392–403.

Andrews, R. L., Ainslie, A., & Currim, I. S. (2002a). An Empirical Comparison of Logit Choice Models with Discrete versus Continuous Representations of Heterogeneity. *Journal of Marketing Research*, 39(4), 479–487.

Andrews, R. L., Ansari, A., & Currim, I. S. (2002b). Hierarchical Bayes versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery. *Journal of Marketing Research*, 39(1), 87–98.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6), 1152–1174.

Aribarg, A., Burson, K. A., & Larrick, R. P. (2017). Tipping the Scale: The Role of Discriminability in Conjoint Analysis. *Journal of Marketing Research*, 54(2), 279–292.

Baumgartner, B., & Steiner, W. J. (2007). Are Consumers Heterogeneous in their Preferences for Odd and Even Prices? Findings from a Choice-Based Conjoint Study. *International Journal of Research in Marketing*, 24(4), 312–323.

Brand Finance (2020). Automotive Industry 2020: The Annual Report on the Most Valuable and Strongest Automobile, Tire, Auto Component & Car Rental Services Brands. *Brand Value Report*, London.

Chen, Y., Iyengar, R., & Iyengar, G. (2017). Modeling Multimodal Continuous Heterogeneity in Conjoint Analysis – A Sparse Learning Approach. *Marketing Science*, 36(1), 140–156.

Conley, T. G., Hansen, C. B., McCulloch, R. E., & Rossi, P. E. (2008). A Semi-Parametric Bayesian Approach to the Instrumental Variable Problem. *Journal of Econometrics*, 144(1), 276–305.

DeSarbo, W. S., Ramaswamy, V., & Cohen, S. H. (1995). Market Segmentation with Choice-Based Conjoint Analysis. *Marketing Letters*, 6(2), 137–147.

Dubé, J.-P., Hitsch, G. J., & Jindal, P. (2014). The Joint Identification of Utility and Discount Functions from Stated Choice Data: An Application to Durable Goods Adoption. *Quantitative Marketing and Economics*, 12(4), 331–377.

Elshiewy, O., Guhl, D., & Boztuğ, Y. (2017). Multinomial Logit Models in Marketing – From Fundamentals to State-of-the-Art. *Marketing ZFP – Journal of Research Management*, 39(3), 32–49.

Escobar, M. D., & West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430), 557–588.

Hauser, J. R. (1978). Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. *Operations Research*, 26(3), 406–421.

Hein, M., Goeken, N., Kurz, P., & Steiner, W. J. (2022). Using Hierarchical Bayes Draws for Improving Shares of Choice Predictions in Conjoint Simulations: A Study based on Conjoint Choice Data. *European Journal of Operational Research*, 297(2), 630–651.

Hein, M., Kurz, P., & Steiner, W. J. (2019). On the Effect of HB Covariance Matrix Prior Settings: A Simulation Study. *Journal of Choice Modelling*, 31, 51–72.

Hein, M., Kurz, P., & Steiner, W. J. (2020). Analyzing the Capabilities of the HB Logit Model for Choice-Based Conjoint Analysis: A Simulation Study. *Journal of Business Economics*, 90(1), 1–36.

Jervis, S. M., Lopetcharat, K., & Drake, M. A. (2012). Application of Ethnography and Conjoint Analysis to Determine Key Consumer Attributes for Latte-Style Coffee Beverages. *Journal of Sensory Studies*, 27(1), 48–58.

Johnson, R., & Orme, B. K. (1996). Getting the Most from CBC. *Sawtooth Software Research Paper Series*, Sequim.

Krueger, R., Vij, A., & Rashidi, T. H. (2018). A Dirichlet Process Mixture Model of Discrete Choice. *Working Paper*, *Cornell University*, *New York*.

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects. *Psychometrika*, 65(1), 93–119.

Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science*, 15(2), 173–191.

Löffler, S., & Baier, D. (2015). Bayesian Conjoint Analysis in Water Park Pricing: A New Approach Taking Varying Part Worths for Attribute Levels into Account. *Journal of Service Science and Management*, 8(1), 46–56.

Louviere, J. J., & Woodworth, G. (1983). Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research*, 20(4), 350–367.

McFadden, D. (1977). Quantitative Methods for Analyzing Travel Behavior of Individuals: Some Recent Developments. *Cowles Foundation Discussion Papers 474. Cowles Foundation for Research in Economics*, *Yale University*.

Moore, W. L. (2004). A Cross-Validity Comparison of Rating-Based and Choice-Based Conjoint Analysis Models. *International Journal of Research in Marketing*, 21(3), 299–312.

Natter, M., & Feurstein, M. (2002). Real World Performance of Choice-Based Conjoint Models. *European Journal of Operational Research*, 137(2), 448–458.

Ogawa, K. (1987). An Approach to Simultaneous Estimation and Segmentation in Conjoint Analysis. *Marketing Science*, 6(1), 66–81.

Ohlssen, D. I., Sharples, L. D., & Spiegelhalter, D. J. (2007). Flexible Random-Effects Models Using Bayesian Semi-Parametric Models: Applications to Institutional Comparisons. *Statistics in Medicine*, 26(9), 2088–2112.

Otter, T., Tüchler, R., & Frühwirth-Schnatter, S. (2004). Capturing Consumer Heterogeneity in Metric Conjoint Analysis Using Bayesian Mixture Models. *International Journal of Research in Marketing*, 21(3), 285–297.

Paetz, F., Hein, M., Kurz, P., & Steiner, W. J. (2019). Latent Class Conjoint Choice Models: A Guide for Model Selection, Estimation, Validation, and Interpretation of Results. *Marketing ZFP – Journal of Research Management*, 41(4), 3–20.

Paetz, F., & Steiner, W. J. (2017). The Benefits of Incorporating Utility Dependencies in Finite Mixture Probit Models. *OR Spectrum*, 39, 793–819.

Ramaswamy, V., & Cohen, S. H. (2007). Latent Class Models for Conjoint Analysis. In A. Gustafsson, A. Hermann, & F. Huber (Eds.), *Conjoint Measurement*, Berlin Heidelberg: Springer, 295–319.

Rossi, P. E. (2014). *Bayesian Non- and Semi-Parametric Methods and Applications*, Princeton: Princeton University Press.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Chichester: John Wiley & Sons.

Sawtooth Software (2017). The CBC System for Choice-Based Conjoint Analysis: Technical Paper (Version 9). *Sawtooth Software Technical Paper Series*, Orem.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2), 639–650.

Silberhorn, N., Boztuğ, Y., & Hildebrandt, L. (2008). Estimation with the Nested Logit Model: Specifications and Software Particularities. *OR Spectrum*, 30, 635–653.

Teichert, T. (2001). Nutzenermittlung in wahlbasierter Conjoint-Analyse: Ein Vergleich von Latent-Class- und hierarchischem Bayes-Verfahren. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, 53, 798–822.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*, 2nd Ed., New York: Cambridge University Press.

Train, K. E., & Sonnier, G. (2005). Mixed Logit with Bounded Distributions of Correlated Partworths. In R. Scarpa, & A. Alberini (Eds.), *Applications of Simulation Methods in Environmental and Resource Economics. The Economics of Non-Market Goods and Resources*, Dordrecht: Springer, 117–134.

van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2002). How Promotions Work: SCAN*PRO-Based Evolutionary Model Building. *Schmalenbach Business Review*, 54, 198–220.

Voleti, S., Srinivasan, V., & Ghosh, P. (2017). An Approach to Improve the Predictive Power of Choice-Based Conjoint Analysis. *International Journal of Research in Marketing*, 34(2), 325–335.

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*, 2nd Ed., Boston: Kluwer Academic Publishers.

Keywords

Choice-based Conjoint Analysis, Preference Heterogeneity, Multinomial Logit Models, Hierarchical Bayesian Estimation, Dirichlet Process Mixture.