Logit mixture with inter and intra-consumer heterogeneity and flexible mixing distributions

Danaf, Mazen; Atasoy, Bilge; Ben-Akiva, Moshe

DOI
10.1016/j.jocm.2019.100188

Publication date
2019

Document Version
Proof

Published in
Journal of Choice Modelling

ARTICLE IN PRESS

Journal of Choice Modelling xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: http://www.elsevier.com/locate/jocm

Logit mixture with inter and intra-consumer heterogeneity and flexible mixing distributions

Mazen Danaf[a,*], Bilge Atasoy[b], Moshe Ben-Akiva[c]

[a] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA
[b] Department of Maritime and Transport Technology, Delft University of Technology, Mekelweg 2, Delft, 2628 CD, the Netherlands
[c] Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Logit mixture
Bayesian estimation
Latent class.
Intra-consumer heterogeneity
Flexible mixing distributions

ABSTRACT

Logit mixture models have gained increasing interest among researchers and practitioners because of their ability to capture unobserved taste heterogeneity. Becker et al. (2018) proposed a Hierarchical Bayes (HB) estimator for logit mixtures with inter- and intra-consumer heterogeneity (defined as taste variations among different individuals and among different choices made by the same individual respectively). However, the underlying model relies on strong assumptions on the inter- and intra-consumer mixing distributions; these distributions are assumed to be normal (or log-normal), and the intra-consumer covariance matrix is assumed to be the same for all individuals. This paper presents a latent class extension to the model and the estimator proposed by Becker et al. (2018) to account for flexible, semi-parametric mixing distributions. This relaxes the normality assumptions and allows different individuals to have different intra-consumer covariance matrices. The proposed model and the HB estimator are validated using real and synthetic data sets, and the models are evaluated using goodness-of-fit statistics and out-of-sample validation. Our results show that when the data comes from two or more distinct classes (with different population means and inter- and intra-consumer covariance matrices), this model results in a better fit and predictions compared to the single class model.

1. Introduction

Because of their ability to capture unobserved taste heterogeneity, logit mixture models have gained increasing interest among researchers and practitioners. These models can be estimated using classical and Bayesian methods, the most common of which is Maximum Simulated Likelihood (MSL). In the Bayesian context, the Hierarchical Bayes (HB) estimator for logit mixture models has been widely applied and documented (Allenby, 1997; Allenby and Rossi, 1998; Train, 2009; etc.). This estimator uses three Gibbs layers, drawing from the population means, covariance matrix, and individual-specific parameters respectively.

Becker et al. (2018) extended this estimator to account for inter- and intra-consumer heterogeneity (defined as taste variations among different individuals and among different choices made by the same individual respectively). However, the underlying model relies on strong assumptions on the inter- and intra-consumer mixing distributions of preferences.

The first assumption is that the mixing distributions are normal (or log-normal), i.e. the choice-specific parameters are normally distributed around the individual-specific means, which are in turn normally distributed around the population means. The key

* Corresponding author.
E-mail addresses: mdanaf@mit.edu (M. Danaf), b.atasoy@tudelft.nl (B. Atasoy), mba@mit.edu (M. Ben-Akiva).

https://doi.org/10.1016/j.jocm.2019.100188
Received 9 April 2019; Received in revised form 29 August 2019; Accepted 20 October 2019
Available online 23 October 2019
1755-5345/

limitation of these parametric distributions is the assumption of uni-modality (Hess, 2014). Several studies have shown that nonparametric mixture models (i.e. latent class) can sometimes outperform continuous mixture models (e.g. Vij and Krueger, 2017), especially when the number of observations per individual is small (Andrews et al., 2002). According to Vij and Krueger (2017), semi-parametric distributions allow for modeling complex patterns of heterogeneity that cannot be captured using uni-modal distributions. These distributions can asymptotically mimic any shape (however, this is not always possible due to data and computational limitations).

The second assumption is that the intra-consumer covariance matrix is assumed to be the same for all individuals, which means that the level of heterogeneity across different choices of a given individual is the same across the population. Becker et al. (2018) justified this assumption because it is not possible to estimate individual-specific covariance matrices given the small number of observations per individual in typical datasets.

This paper extends the logit double mixture model and the estimator proposed by Becker et al. (2018) to account for flexible, semi-parametric mixing distributions of unobserved taste heterogeneity. This extension is an important contribution as it overcomes the abovementioned limitations:

1. It relaxes the normality assumptions on the mixing distribution of inter-consumer heterogeneity, which allows for approximating any shape. For example, a distribution with S modes can be approximated by a normal mixture of S classes.
2. It allows different individuals to have different intra-consumer covariance matrices, depending on the class to which they belong.

Semi-parametric distributions have not been used before in logit mixture models with inter- and intra-consumer heterogeneity, even though they have proven useful in models with inter-consumer heterogeneity only (Rossi et al., 2005; Bujosa et al., 2010; Greene and Hensher, 2013; Keane and Wasi, 2013; and Krueger et al., 2018). In this paper, we propose a Hierarchical Bayes (HB) estimator for these models based on Gibbs sampling with embedded Metropolis-Hastings (MH) algorithms, and demonstrate that they can outperform logit mixture models with unimodal (normal or log-normal) distributions of inter- and intra-consumer heterogeneity. Finally, we show that the optimal number of classes can be determined empirically using out-of-sample validation.

The remainder of the paper is divided as follows. Section 2 presents a brief background on flexible mixing distributions and models with intra-consumer heterogeneity. Section 3 presents the proposed extension and the Hierarchical Bayes (HB) estimator. Section 4 presents an application with synthetic data to validate the estimator presented in section 3. Another application with real data is presented in Section 5. Finally, Section 6 presents a discussion of the applications, advantages, and limitations of the model, and Section 7 concludes the paper.

2. Background

This section presents a brief overview of flexible mixing distributions in choice models, and choice models with inter- and intra-consumer heterogeneity.

2.1. Flexible mixing distributions

The majority of studies in discrete choice models have used normal and log-normal distributions, with a few using Johnson's S_b gamma, and triangular distributions (Train, 2016). Because of the limitations associated with these distributions, several extensions have been proposed in the literature including latent class models, models that account for scale heterogeneity, and discrete mixtures of continuous distributions (also known as mixed-mixed logit models).

Latent class models are semi-parametric models that assume a finite number of market segments, with preferences being homogenous within each segment. These models are advantageous as they do not require numerical simulation. However, they have been found to understate the extent of heterogeneity in the data (Elrod and Keane, 1995; Allenby and Rossi, 1998).

On the other hand, several procedures have been proposed for modeling continuous flexible mixing distributions (Bajari et al., 2007; Fosgerau and Bierlaire, 2007; Burda et al., 2008; Train, 2008; Fox et al., 2011; Train, 2016, etc.). The most popular procedure is the - semi-parametric - mixture of normals, also known as the mixed-mixed logit/probit model. Such models have been proposed for logit (Rossi et al., 2005; Bujosa et al., 2010; Greene and Hensher, 2013; Keane and Wasi, 2013; and Krueger et al., 2018) and probit models (Geweke and Keane, 2001, 2007). These models represent a combination of continuous mixture models and discrete mixture/latent class models, and are convenient for estimation using MSL and HB (Rossi et al., 2005; Sarrias and Daziano, 2017).

2.2. Inter- and intra-consumer heterogeneity

The abovementioned studies have focused on modeling inter-consumer heterogeneity using flexible mixing distributions. However, in the presence of multiple observations from each individual, it is possible to identify inter as well as intra-consumer heterogeneity. Models with intra-consumer heterogeneity have been estimated using MSL (Bhat and Castelar, 2002; Bhat and Sardesai, 2006; Hess and Rose, 2009; Hess and Train, 2011; and Yáñez et al., 2011), HB (Becker et al., 2018; Ben-Akiva et al., 2019), and maximum approximate composite marginal likelihood (MACML) (Bhat and Sidharthan, 2011).

Hess and Train (2011) argue that when the data contains multiple choices by each consumer, it is natural to assume that preferences vary across choice situations for the same individual. They show that intra-consumer heterogeneity can be estimated for panel data and achieves a better fit compared to simpler models (i.e. those with only inter-consumer heterogeneity). Hess and Rose (2009) also justify modeling inter-as well as intra-consumer heterogeneity for reasons such as non-linearities in response, learning and fatigue effects, thresholds, and variations in scale across choice situations.

According to Ben-Akiva et al. (2019), ignoring intra-consumer heterogeneity assumes a nearly neoclassical consumer with "permanent" individual preferences, and treats perturbations in these preferences as nuisance factors. Becker et al. (2018) and Ben-Akiva et al. (2019) also show that falsely ignoring intra-consumer heterogeneity despite its presence in the data leads to biased estimates, inflated scales, and a decreased goodness-of-fit.

The main limitation of estimating models with inter- and intra-consumer heterogeneity is the excessively long estimation time. According to Bhat and Sidharthan (2011), MSL estimation of these models is practically infeasible when the mixing structure leads to an explosion in the dimensionality of integration in the likelihood function.

HB and MACML methods have been proposed as alternatives to MSL to enable the estimation of choice models with inter- and intra-consumer heterogeneity. Bhat (2011) introduced the MACML estimator of the standard multinomial probit (MNP) mixture model (with only inter-consumer heterogeneity), and proposed an extension for models with inter- and intra-consumer heterogeneity. This estimator was used by Bhat and Sidharthan (2011), who showed that in a Monte Carlo simulation, MACML was about 350 times or more faster than MSL for panel data with inter- and intra-consumer heterogeneity using a MNP mixture model. Similarly, Becker et al. (2018) showed that for multinomial logit mixture models, HB outperforms MSL significantly in terms of run-time; for a simple Monte Carlo experiment with only 500 individuals and four parameters (including a scale parameter and one random parameter with inter- and intra-consumer heterogeneity), estimation time using MSL takes about 11 h with only 500 draws, and 377 h with 1 000 draws, while HB estimation takes about 2 h (Becker et al. (2018) used numerical gradients, however, the MSL run time can be substantially reduced by using analytical gradients).

2.3. The HB estimator for logit mixtures with inter- and intra-consumer heterogeneity

Becker et al. (2018) and Ben-Akiva et al. (2019) proposed an HB estimator for logit models with inter- and intra-consumer heterogeneity that overcomes the computational constraints associated with MSL estimators. The authors consider the "double mixture" model with three levels of parameters:

1. Population-level parameters μ and Ω^b: represent the average tastes/preferences in the population and the inter-consumer covariance matrix respectively.
2. Individual-level parameters ζ_n and Ω^w: represent the average tastes/preferences of a specific individual and the intra-consumer covariance matrix respectively.
3. Choice-specific parameters η_{mn}: reflect the tastes/preferences specific to each choice situation.

where n is an index for individuals ($n = 1, 2, \ldots, N$), m is an index of choice situations ($m = 1, 2, \ldots, M_n$), and j is an index for alternatives ($j = 1, 2, \ldots, J_{mn}$).

The probability of individual n choosing alternative j in choice m is shown in equation (1):

$$P\left(d_{jmn} = 1|\mu, \Omega^b, \Omega^w\right) = \int_{\zeta_n}\int_{\eta_{mn}} P_j(\eta_{mn})\ \ H\left(d\eta_{mn}|\zeta_n, \Omega^w\right)F\left(d\zeta_n|\mu, \Omega^b\right) \tag{1}$$

where d_{jmn} is equal to one if individual n chooses alternative j in choice m and zero otherwise, and:

$$P_j(\eta_{mn}) = \frac{\exp\left(V_j(\eta_{mn})\right)}{\sum_{j'=1}^{J_{mn}}\exp\left(V_{j'}(\eta_{mn})\right)} \tag{2}$$

$$H\left(d\eta_{mn}|\zeta_n, \Omega^w\right) \sim \mathscr{N}_T(\zeta_n, \Omega^w) \tag{3}$$

$$F\left(d\zeta_n|\mu, \Omega^b\right) \sim \mathscr{N}_T\left(\mu, \Omega^b\right) \tag{4}$$

where T is the number of unknown parameters and $\mathscr{N}_T$ represents the T-dimensional multivariate normal distribution.

The HB estimator extends the standard estimator of the logit mixture model (Allenby, 1997; Train, 2009) by adding two additional Gibbs layers sampling from choice-specific parameters (η_{mn}) and the intra-consumer covariance matrix (Ω_w). The Gibbs sampler can be summarized as follows:

1. $\mu|\Omega^b, \Omega^w, \zeta_n, \eta_{mn}$: Normal Bayesian update with unknown mean and known variance, using a diffuse prior and ζ_n as the data.
2. $\Omega^b|\mu, \Omega^w, \zeta_n, \eta_{mn}$: Normal Bayesian update with known mean and unknown variance, using a diffuse prior and ζ_n as the data.
3. $\Omega^w|\mu, \Omega^b, \zeta_n, \eta_{mn}$: Normal Bayesian update with known mean and unknown variance, using a diffuse prior and $\eta_{mn} - \zeta_n$ as the data.
4. $\zeta_n|\mu, \Omega^b, \Omega^w, \eta_{mn}$: Normal Bayesian update with unknown mean and known variance, using η_{mn} as the data. The density of the population parameters serves as the prior for each individual-specific parameter.
5. $\eta_{mn}|\mu, \Omega^b, \Omega^w, \zeta_n$: Metropolis-Hastings procedure, where the conditional posterior is proportional to Logit multiplied by a normal density.

In the following section, we present an extension to the logit double mixture model using semi-parametric distributions. The population is divided into two or more segments, each characterized by its inter- and intra-consumer mixing distributions. The distributions of inter- and intra-consumer heterogeneity are assumed to be normally distributed within each segment, which maintains the desirable properties of the normal distribution (i.e. conjugacy) in the Gibbs sampler.

3. Proposed extension

In this section, we build on the model and the HB estimator proposed by Becker et al. (2018) and Ben-Akiva et al. (2019) to model inter-as well as intra-consumer heterogeneity using semi-parametric distributions.

3.1. The model

The model extends the logit double mixture model by assuming that the population is composed of S classes/segments where $S \geq 2$. Each class has its own population means (μ_s) and inter- and intra-consumer covariance matrices (Ω_s^b, Ω_s^w). We also define s_n as a class membership indicator for each individual:

$$s_n \in \{\mathbb{Z} : [1, S]\} \tag{5}$$

We assume that ζ_n and η_{mn} are normally distributed within each class:

$$\zeta_{n|s_n=s} \sim \mathcal{N}_T(\mu_s, \Omega_s^b) \tag{6}$$

$$\eta_{mn|s_n=s} \sim \mathcal{N}_T(\zeta_n, \Omega_s^w) \tag{7}$$

Finally, we define the marginal probability of belonging to class s as π_s, and π as the vector $(\pi_1, \ldots, \pi_s)$:

$$\pi_s = P(s_n = s) \tag{8}$$

Conditional on the class membership, the probability of individual n choosing alternative j in choice situation m is:

$$\begin{aligned} &P(d_{jmn} = 1 | s_n = s) = P(d_{jmn} = 1 | \mu_s, \Omega_s^b, \Omega_s^w) \\ &= \int_{\zeta_n} \int_{\eta_{mn}} P_j(\eta_{mn}) \; H(d\eta_{mn} | \zeta_n, \Omega_s^w) F(d\zeta_n | \mu_s, \Omega_s^b) \end{aligned} \tag{9}$$

Where:

$$H(d\eta_{mn} | \zeta_n, \Omega_s^w) \sim \mathcal{N}_T(\zeta_n, \Omega_s^w) \tag{10}$$

$$F(d\zeta_n | \mu_s, \Omega_s^b) \sim \mathcal{N}_T(\mu_s, \Omega_s^b) \tag{11}$$

Since the class membership is unknown, the unconditional probability is obtained as a weighted sum over all the possible classes (weighted by π_s):

$$\begin{aligned} P(d_{jmn} = 1 | \mu_s, \Omega_s^b, \Omega_s^w, \pi_s \;\; \forall \;\; s \in \{\mathbb{Z} : [1, S]\}) &= \sum_{s=1}^{S} \pi_s P\left(d_{jmn} = 1 | \mu_s, \Omega_s^b, \Omega_s^w\right) \\ &= \sum_{s=1}^{S} \pi_s \int_{\zeta_n} \int_{\eta_{mn}} P_j(\eta_{mn}) \; H(d\eta_{mn} | \zeta_n, \Omega_s^w) F(d\zeta_n | \mu_s, \Omega_s^b) \end{aligned} \tag{12}$$

3.2. Model estimation

This model can be estimated by extending the 5-steps Gibbs sampler as follows. Hierarchical Bayes (HB) and data augmentation are needed, where we assume that s_n, ζ_n, and η_{mn} are unknown parameters. We also use a diffuse prior on π denoted by $k(\pi)$.

The posterior on ζ_n, s_n, η_{mn}, μ_s, Ω_s^w, $\Omega_s^b \; \forall \, s$, and π is given by equation (13):

$$\begin{aligned} &K(\zeta_n, s_n \;\; \forall n, \eta_{mn} \;\; \forall mn, \mu_s, \Omega_s^w, \Omega_s^b, \pi_s \;\; \forall \;\; s | d) \propto \\ &\prod_{n=1}^{N} \left[\prod_{m=1}^{M_n} \left[\prod_{j=1}^{J_{mn}} \left[P_j(\eta_{mn})^{d_{jmn}} \right] h(\eta_{mn} | \zeta_n, \Omega_{s_n}^w) \right] f(\zeta_n | \mu_{s_n}, \Omega_{s_n}^b) P(s_n | \pi) \right] k(\pi) \prod_{s=1}^{S} k(\Omega_s^w) k(\mu_s) k(\Omega_s^b) \end{aligned} \tag{13}$$

Where:

$$k(\mu_s) \sim \mathcal{N}_T(\mu_{0,s}, \;\; A_s) \tag{14}$$

$$k(\Omega_s^b) \sim HIW(\nu_{bs}, A_{bs}) \tag{15}$$

$$k(\Omega_s^w) \sim HIW(\nu_{ws}, A_{ws}) \tag{16}$$

$$k(\pi) \sim DIR(\alpha) \tag{17}$$

where $\mu_{0,s}$ represents a vector of prior means, A_s is a diagonal covariance matrix with diagonal values → ∞ (uninformative prior), HIW is the Hierarchical Inverted Wishart (Half-t) prior with parameters ν_{bs}, ν_{ws}, A_{bs}, and A_{ws} (Huang and Wand, 2013; Akinc and Vandebroek, 2018), and α is a vector of concentration parameters (e.g. $\alpha_s = 1 \ \forall \ s$). The HIW (Half-t) prior is recommended by Huang and Wand (2013) and Akinc and Vandebroek (2018), who showed that the latter outperforms the commonly used Inverted Wishart (IW) priors in terms of noninformativity and parameter recovery.

This model is estimated using a 7-step Gibbs sampler which draws from the conditional posteriors below:

Step I: drawing from the conditional posterior of the population means for each class:

$$K(\mu_s | \zeta_n \ , s_n \forall n, \eta_{mn} \ \ \forall mn, \Omega_s^w, \Omega_s^b, \pi_s) \propto f(\zeta_n \forall n_{|s_n = s} | \mu_s, \Omega_s^b) k(\mu_s) \tag{18}$$

This includes S Normal Bayesian updates with known variances and unknown means. The conditional posterior on each μ_s is $\mathcal{N}\left(\bar{\zeta}_s, \frac{\Omega_s^b}{N_s}\right)$, $\bar{\zeta}_s$ is the mean of ζ_n draws over all individuals who belong to class s in the previous Gibbs iteration ($s_n = s$), and N_s is the number of individuals belonging to class s in previous iteration.

Step II: drawing from the conditional posterior of the inter-consumer covariance matrix for each class:

$$K(\Omega_s^b | \zeta_n, s_n \ \ \forall n, \eta_{mn} \ \ \forall mn, \mu_s, \Omega_s^w, \pi_s) \propto f(\zeta_n \forall n_{|s_n = s} | \mu_s, \Omega_s^b) k(\Omega_s^b) \tag{19}$$

This includes S normal Bayesian updates with unknown variances and known means, similar to the original HB estimator. However, only observations belonging to class s in the previous iteration are used in updating the inter-consumer covariance matrix of the corresponding class.

Step III: drawing from the conditional posterior of the class probabilities π_s:

$$K(\pi_1, \ldots, \pi_s | \zeta_n, s_n \ \ \forall n, \eta_{mn} \ \ \forall mn, \mu_s, \Omega_s^w, \Omega_s^b) \propto P(s_n | \pi) k(\pi) \tag{20}$$

This step draws from a Dirichlet distribution with parameters $N_s + \alpha_s$ for each class s. Alternatively, a class membership model can be specified as a function of individual characteristics, and this step can be replaced by a Metropolis-Hastings (MH) algorithm. This extension is discussed in Section 3.3.

Step IV: drawing from the conditional posterior of the class membership indicators $s_n \ \forall \ n$:

$$P(s_n | \mu_s, \Omega_s^b, \Omega_s^w \ \forall \ s, \pi_s, \zeta_n \ \forall \ n, \ \eta_{mn} \forall \ mn) \propto P(s_n | \pi) f(\zeta_n | \mu_{s_n}, \Omega_{s_n}^b) \prod_{m=1}^{M_n} h(\eta_{mn} | \zeta_n, \Omega_{s_n}^w) \tag{21}$$

This can be performed by calculating the class probabilities for each individual and simulating the class membership:

$$P(s_n = s | \mu_s, \Omega_s^b, \Omega_s^w, \pi_s \forall s, \ \zeta_n \ \forall \ n, \ \eta_{mn} \forall \ mn) = \frac{\pi_s f(\zeta_n | \mu_s, \Omega_s^b) \prod_{m=1}^{M_n} h(\eta_{mn} | \zeta_n, \Omega_s^w)}{\sum_{k=1}^{S} \left[\pi_k f(\zeta_n | \mu_k, \Omega_k^b) \prod_{m=1}^{M_n} h(\eta_{mn} | \zeta_n, \Omega_k^w)\right]} \tag{22}$$

If individual characteristics are used to model class membership, the fitted probability of class s for each individual is used in equation (22) instead of π_s.

Step V: drawing from the conditional posterior of the intra-consumer covariance matrix for each class:

$$K(\Omega_s^w | s_n, \zeta_n \ \ \forall n, \eta_{mn} \ \ \forall mn, \mu_s, \Omega_s^b, \pi_s) \propto h(\eta_{mn} \ \ \forall mn_{|s_n = s} | \zeta_n \ \ \forall n, \Omega_s^w) k(\Omega_s^w) \tag{23}$$

This includes S normal Bayesian updates with unknown variances and known means. However, only observations belonging to class s in this iteration are used in updating the intra-consumer covariance matrix of the corresponding class.

Step VI: drawing from the conditional posterior of the individual-level means:

$$K(\zeta_n | \mu_s, \eta_{mn}, \Omega_s^b, \Omega_s^w, s_n, \pi) \propto h(\eta_{mn} | \zeta_n \ \ \forall n, \Omega_{s_n}^w) f(\zeta_n | \mu_{s_n}, \Omega_{s_n}^b) \tag{24}$$

This is performed using a normal Bayesian update with η_{mn} as the data and the distribution $\mathcal{N}(\mu_{s_n}, \Omega_{s_n}^b)$ as a prior.

Step VII: drawing from the conditional posterior of the individual- and choice-specific coefficients:

$$K\left(\eta_{mn}|\mu_s,\zeta_n,\Omega_s^b,\Omega_s^b,\pi,s_n\right)\propto\prod_{j=0}^{J_{mn}}\left[P_j(\eta_{mn})^{d_{jmn}}\right]h\left(\eta_{mn}|\zeta_n,\Omega_{s_n}^w\right)$$

$$n=1,\ 2,\ \ldots,\ N,\ \ m=1,\ \ldots M_n \tag{25}$$

A draw of η_{mn} is obtained by the MH algorithm, where the jumping distribution is $\mathcal{N}(\zeta_n,\Omega_{s_n}^w)$.

3.3. Class membership models

Class membership can also be modeled as a function of individual characteristics and socio-economic variables. Instead of using marginal class probabilities (π_s), a class membership model ($Q_s(\theta)$) can be used (parametrized with θ). For example, a logit model can be specified to model the probability of belonging to class s as a function of socio-demographic variables and the parameters θ. The modified model is shown in equation (26).

$$P\left(d_{jmn}=1|\mu_s,\Omega_s^b,\Omega_s^w,\theta\ \ \forall\ \ s\in\{\mathbb{Z}:[1,S]\right)=$$

$$\sum_{s=1}^{S}Q_s(\theta)\int_{\zeta_n}\int_{\eta_{mn}}P_j(\eta_{mn})\ \ H\left(d\eta_{mn}|\zeta_n,\Omega_s^w\right)F\left(d\zeta_n|\mu_s,\Omega_s^b\right) \tag{26}$$

In estimation, the individual-specific class membership indicators obtained from step 4 (s_n) are used as dependent variables in the logit model, and Step 3 is replaced with a Metropolis-Hastings algorithm drawing from the class membership model parameters (θ). This extension is used in the following applications in Sections 4 and 5.

4. Experiments with synthetic data

4.1. Synthetic data description

Synthetic data is used in order to validate the estimator presented above. The data mimics a stated preferences (SP) experiment in which individuals are presented with Mobility-*as*-a-Service (MaaS) plans. We assume that each individual is presented with 8 menus (choice situations), and that each menu includes three MaaS plans. Each plan has two binary attributes (unlimited access to transit (T) and unlimited access to bike sharing (B)), and two non-binary attributes (price (P) and the number of on-demand trips per month (OD)).

The systematic utility equations are given by equation (27):

$$\begin{aligned}V_{1mn}&=\exp(\alpha_{mn})(-P_{1mn}+\beta_{T,mn}T_{1mn}+\beta_{B,mn}B_{1mn}+\exp(\beta_{OD,mn})OD_{1mn}+C_1)\\V_{2mn}&=\exp(\alpha_{mn})(-P_{2mn}+\beta_{T,mn}T_{2mn}+\beta_{B,mn}B_{2mn}+\exp(\beta_{OD,mn})OD_{2mn}+C_2)\\V_{3mn}&=\exp(\alpha_{mn})(-P_{3mn}+\beta_{T,mn}T_{3mn}+\beta_{B,mn}B_{3mn}+\exp(\beta_{OD,mn})OD_{3mn})\end{aligned} \tag{27}$$

Where:

- P_{jmn} is the price of plan j in menu m (ranging between 10 and 160 USD)

Table 1
True values of the parameters in the synthetic data.

Class 1			
Parameter	Population mean	Inter Consumer Var.	Intra Consumer Var.
Scale	0.50	0.25	0.10
Transit	4.00	1.00	0.25
Bike Sharing	3.00	1.00	0.25
On-Demand	−2.50	0.50	0.25
Class 2			
	Population mean	Inter Consumer Var.	Intra Consumer Var.
Scale	0.50	0.25	0.10
Transit	1.00	0.50	0.50
Bike Sharing	1.00	0.50	0.50
On-Demand	−1.50	0.25	0.50
Constant Parameters			
Constant - Alternative 1	1.00		
Constant - Alternative 2	0.50		
Constant - Class 1	−0.10		
Student - Class 1	1.00		

- T_{jmn} and B_{jmn} are binary variables representing unlimited access to transit and bike-sharing respectively.
- D_{jmn} represents the number of on-demand trips per month (between 5 and 30 trips).
- C_1 and C_2 are non-random alternative specific constants for choosing alternatives 1 and 2 respectively (the constant for alternative 3 is normalized to 0).
- α_{mn} is a scale parameter distributed with inter- and intra-consumer heterogeneity. Exponentiation is used in order to ensure that it is positive for all individuals and menus.
- $\beta_{T,mn}$, $\beta_{B,mn}$, and $\beta_{OD,mn}$ are the parameters of transit, bike sharing, and on-demand respectively.

We assume that the population consists of two classes, and that the parameters ($\beta_{T,mn}$, $\beta_{B,mn}$, and $\beta_{OD,mn}$) are normally distributed within each class with inter- and intra-consumer heterogeneity.

The true values of the model parameters are presented in Table 1. Class 1 has a higher preference to transit and bike sharing access, while Class 2 has a higher preference to on-demand trips. In addition, Class 1 has high inter-consumer heterogeneity and low intra-consumer heterogeneity, while Class 2 has high intra-consumer heterogeneity and low inter-consumer heterogeneity. The inter- and intra-consumer distributions of the scale parameter are the same for the two classes.

Within each class, we assume that there is a positive correlation between the individual-specific means of transit and bike sharing, and negative correlations between the individual- and menu-specific means of on-demand and transit, and on-demand and bike sharing respectively. The inter-and intra-consumer covariances for the two classes are shown in Tables 2 and 3.

Finally, we assume that the probability of belonging to Class 1 is a function of student status (binary). We define $student_n$ as a dummy variable equal to 1 if individual n is a student and 0 otherwise. A binary logit model is used to determine class membership using equation (28).

$$\begin{aligned} V_{class1,n} &= \alpha_0 + \alpha_1 Student_n \\ V_{class2,n} &= 0 \end{aligned} \quad (28)$$

The choices are simulated by calculating the systematic utility of each alternative, and adding an EV (0,1) error term to obtain the total random utility. After accounting for inter- and intra-consumer heterogeneity, the effect of the unobserved error term is relatively small compared to that of the systematic part of the utility (the choices simulated using the systematic utility only match those simulated using the total random utility 86% of the time).

Fig. 1 shows the distributions of the individual-specific means and the choice-specific parameters for the transit, bike sharing, on-demand, and scale parameters. The distributions of the transit parameter are bimodal, while the distributions of the bike sharing and on-demand parameters are unimodal and skewed. The distributions of the scale parameter are symmetric and unimodal as expected (because these distributions are the same for the two classes).

4.2. Model estimation

The sample used in the baseline experiment consists of 5 000 individuals (additional estimations with 2 000 and 20,000 individuals are presented in Section 4.3). Convergence is reached after 400,000 Gibbs iterations, the first 200,000 of which are burn-in iterations, and the remaining 200,000 are used for sampling from the posterior distributions. Convergence is tested using Gelman and Rubin's (1992) potential reduction factor (R-hat). Estimation is repeated twice with different starting values, and 1 000 draws are stored for each parameter (with equal intervals). The R-hat statistic is calculated as a weighted average of the within-chain and between-chain variances as shown in Equation (29). A value close to 1.0 indicates stationarity of the corresponding chain.

$$\widehat{R} = \sqrt{\frac{D-1}{D} W + \frac{1}{D} B} \quad (29)$$

where D is the number of draws, W is the within-chain variance, and B is the between-chain variance.

The estimation results are presented in Table 4, showing the true values, posterior means, posterior standard deviations (in parenthesis), and R-hat statistics. The results indicate that the estimator is able to recover the true values of the model parameters. All of the estimated parameters are within a reasonable range of their true values, and most of them are not statistically different from these true values as indicated by the posterior standard deviations (presented in parentheses). We are also able to estimate the co-variances (shown in Tables 5 and 6) of the inter- and intra-consumer distributions. However, these are calculated with large posterior

Table 2
Off-diagonal elements of the inter-consumer distribution (Class 1 co-variances in the lower triangular part and Class 2 covariances in the upper triangular part).

	Transit	Bike Sharing	On-Demand
Transit		0.20	-0.20
Bike Sharing	0.40		-0.20
On-Demand	-0.40	-0.40	

Table 3
Off-diagonal elements of the intra-consumer distribution (Class 1 co-variances in the lower triangular part and Class 2 covariances in the upper triangular part).

	Transit	Bike Sharing	On-Demand
Transit		0.30	-0.20
Bike Sharing	0.15		-0.20
On-Demand	-0.15	-0.15	

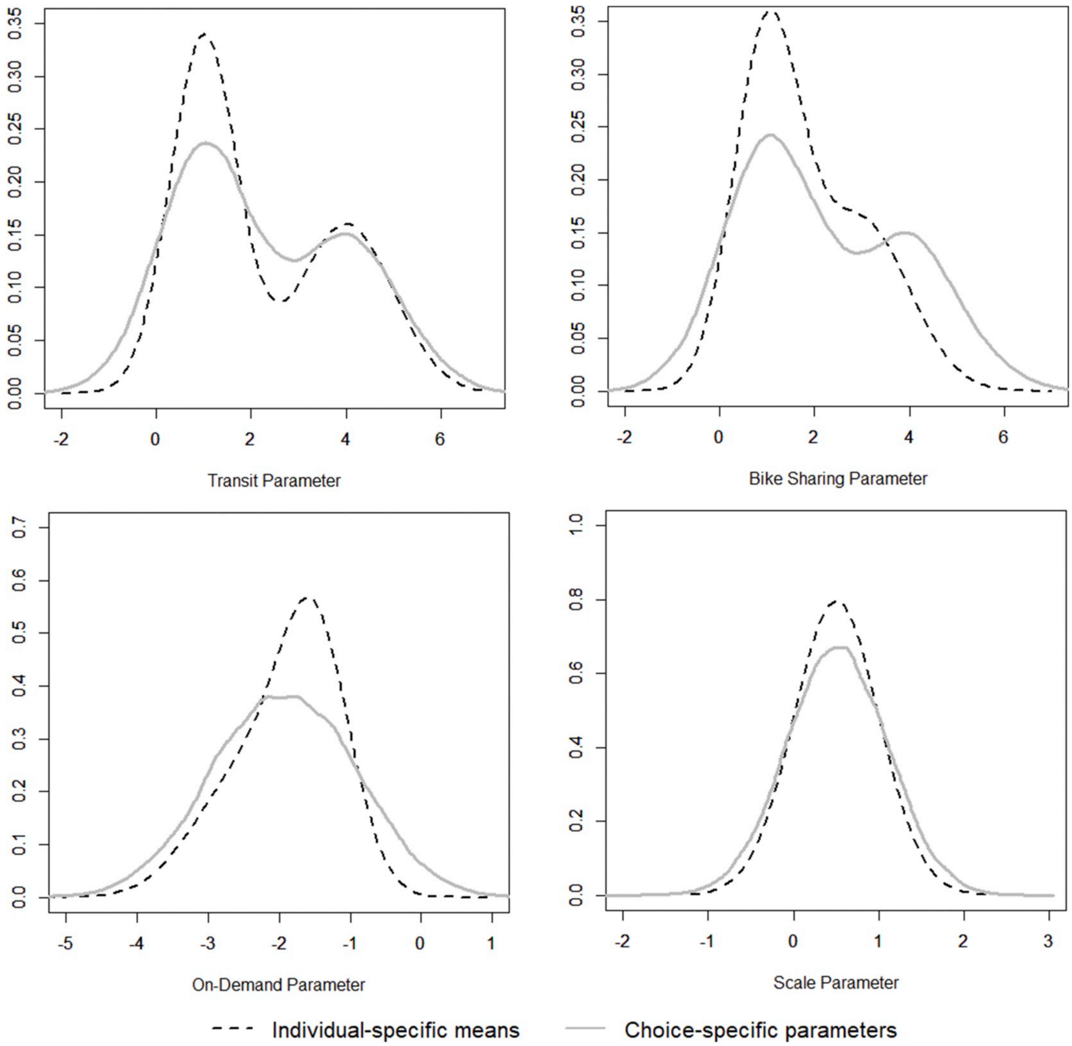

Fig. 1. Inter- and intra-consumer distributions in the synthetic data.

standard deviations. The R-hat values indicate that convergence has been reached, and that 200,000 iterations were enough for estimating this model.

4.3. Varying sample size

In order to further validate the estimator, we rerun estimation with different sample sizes: 2 000 and 2,0000 individuals. The number of observations per individual is fixed at 8. The results (presented in Table 7) show that the posterior means get closer to their

Table 4
Baseline estimation with 5 000 individuals.

Parameter	Class 1 Parameters			Class 2 Parameters		
	True value	Estimate	R-hat	True value	Estimate	R-hat
Transit - population mean	4.00	4.227 (0.116)	1.00	1.00	1.035 (0.075)	1.07
Bike Sharing - population mean	3.00	3.016 (0.096)	1.00	1.00	1.085 (0.050)	1.00
On-Demand - population mean	−2.50	−2.578 (0.07)	1.01	−1.50	−1.526 (0.025)	1.01
Transit - inter-consumer var.	1.00	1.020 (0.210)	1.03	0.50	0.668 (0.128)	1.03
Bike Sharing - inter-consumer var.	1.00	0.755 (0.167)	1.00	0.50	0.693 (0.098)	1.03
On-Demand - inter-consumer var.	0.50	0.519 (0.073)	1.21	0.25	0.252 (0.025)	1.01
Transit - intra-consumer var.	0.25	0.289 (0.148)	1.00	0.50	0.430 (0.118)	1.12
Bike Sharing - intra-consumer var.	0.25	0.317 (0.141)	1.18	0.50	0.356 (0.089)	1.03
On-Demand - intra-consumer var.	0.25	0.236 (0.062)	1.07	0.50	0.514 (0.034)	1.01
Shared Parameters						
	True value		**Estimate**		**R-hat**	
Scale - population mean	0.50		0.477 (0.03)		1.07	
Scale - inter-consumer var.	0.25		0.274 (0.038)		1.02	
Scale - intra-consumer var.	0.10		0.09 (0.04)		1.20	
Constant – Alt. 1	1.00		1.021 (0.02)		1.00	
Constant – Alt. 2	0.50		0.513 (0.02)		1.01	
Constant - Class 1	−0.10		0.073 (0.117)		1.00	
Student - Class 1	1.00		1.093 (0.087)		1.01	

Table 5
Estimated off-diagonal elements of the inter-consumer distribution (Class 1 covariances in the lower triangular part and Class 2 covariances in the upper triangular part).

	Estimated Values			True Values		
	Transit	Bike Sharing	On-Demand	Transit	Bike Sharing	On-Demand
Transit		0.341 (0.083)	-0.241 (0.040)		0.20	-0.20
Bike Sharing	0.532 (0.159)		-0.248 (0.040)	0.40		-0.20
On-Demand	-0.338 (0.112)	-0.36 (0.118)		-0.40	-0.40	

Table 6
Estimated off-diagonal elements of the intra-consumer distribution (Class 1 covariances in the lower triangular part and Class 2 covariances in the upper triangular part).

	Estimated Values			True Values		
	Transit	Bike Sharing	On-Demand	Transit	Bike Sharing	On-Demand
Transit		0.183 (0.076)	-0.201 (0.050)		0.30	-0.20
Bike Sharing	0.134 (0.095)		-0.250 (0.049)	0.15		-0.20
On-Demand	-0.164 (0.082)	-0.077 (0.058)		-0.15	-0.15	

true values as the sample size increases, and the posterior standard deviations decrease as expected. In addition, the variances and covariances are recovered more precisely when the sample size has 20,000 individuals. We also observe that the required number of Gibbs iterations to reach stationarity decreases as the sample size increases.

4.4. Monte Carlo analysis

In this section, we repeat the estimation described in Section 4.2 with 30 different samples obtained from the same data generation process, each consisting of 5 000 individuals. We then compare the distribution of the posterior means obtained from the 30 samples to the true values of the model parameters, and the posterior distributions obtained from one replication respectively. The results are presented in Table 8. These results indicate that the averages of the posterior means over all 30 runs are very close to the true values, and that the standard deviations over all 30 runs are comparable to the posterior standard deviations obtained from a single run.

4.5. Model selection

In this section, we compare the predictions and goodness-of-fit of the estimated model (labelled “*Latent Class Extension*”) to the

Table 7
Estimation with varying sample sizes.

	Parameter	True value	N = 2 000	N = 5 000	N = 20,000
Class 1 Parameters	Transit - population mean	4.00	4.008 (0.261)	4.227 (0.116)	4.071 (0.069)
	Bike Sharing - population mean	3.00	3.019 (0.178)	3.016 (0.096)	3.039 (0.049)
	On-Demand - population mean	−2.50	−2.396 (0.093)	−2.578 (0.070)	−2.570 (0.033)
	Transit - inter-consumer var.	1.00	1.238 (0.457)	1.020 (0.210)	0.905 (0.129)
	Bike Sharing - inter-consumer var.	1.00	0.985 (0.262)	0.755 (0.167)	0.979 (0.063)
	On-Demand - inter-consumer var.	0.50	0.498 (0.087)	0.519 (0.073)	0.526 (0.041)
	Transit - intra-consumer var.	0.25	0.348 (0.189)	0.289 (0.148)	0.307 (0.039)
	Bike Sharing - intra-consumer var.	0.25	0.400 (0.138)	0.317 (0.141)	0.200 (0.067)
	On-Demand - intra-consumer var.	0.25	0.239 (0.083)	0.236 (0.062)	0.283 (0.029)
	Transit/Bike Sharing cov. (inter)	0.40	0.708 (0.271)	0.532 (0.159)	0.347 (0.077)
	Transit/On-Demand cov. (inter)	−0.4	−0.570 (0.156)	−0.338 (0.112)	−0.333 (0.064)
	Bike Sharing/On-Demand cov. (inter)	−0.4	−0.515 (0.114)	−0.360 (0.118)	−0.390 (0.039)
	Transit/Bike Sharing cov. (intra)	0.15	0.126 (0.134)	0.134 (0.095)	0.117 (0.024)
	Transit/On-Demand cov. (intra)	−0.15	−0.117 (0.092)	−0.164 (0.082)	−0.148 (0.026)
	Bike Sharing/On-Demand cov. (intra)	−0.15	−0.113 (0.095)	−0.077 (0.058)	−0.144 (0.021)
Class 2 Parameters	Transit - population mean	1.00	0.993 (0.133)	1.035 (0.075)	1.078 (0.038)
	Bike Sharing - population mean	1.00	1.026 (0.093)	1.085 (0.050)	1.016 (0.036)
	On-Demand - population mean	−1.50	−1.530 (0.040)	−1.526 (0.025)	−1.497 (0.013)
	Transit - inter-consumer var.	0.50	0.643 (0.198)	0.668 (0.128)	0.592 (0.047)
	Bike Sharing - inter-consumer var.	0.50	0.592 (0.124)	0.693 (0.098)	0.527 (0.043)
	On-Demand - inter-consumer var.	0.25	0.232 (0.032)	0.252 (0.025)	0.245 (0.011)
	Transit - intra-consumer var.	0.50	0.725 (0.260)	0.430 (0.118)	0.432 (0.033)
	Bike Sharing - intra-consumer var.	0.50	0.321 (0.095)	0.356 (0.089)	0.543 (0.044)
	On-Demand - intra-consumer var.	0.50	0.443 (0.052)	0.514 (0.034)	0.490 (0.016)
	Transit/Bike Sharing cov. (inter)	0.2	0.207 (0.150)	0.341 (0.083)	0.258 (0.051)
	Transit/On-Demand cov. (inter)	−0.2	−0.195 (0.063)	−0.241 (0.040)	−0.204 (0.019)
	Bike Sharing/On-Demand cov. (inter)	−0.2	−0.177 (0.054)	−0.248 (0.040)	−0.193 (0.019)
	Transit/Bike Sharing cov. (intra)	0.3	0.171 (0.146)	0.183 (0.076)	0.289 (0.044)
	Transit/On-Demand cov. (intra)	−0.2	−0.206 (0.079)	−0.201 (0.050)	−0.155 (0.040)
	Bike Sharing/On-Demand cov. (intra)	−0.2	−0.076 (0.072)	−0.250 (0.049)	−0.205 (0.030)
Shared Parameters	Scale - population mean	0.50	0.461 (0.047)	0.477 (0.030)	0.513 (0.013)
	Scale - inter-consumer var.	0.25	0.233 (0.053)	0.274 (0.038)	0.280 (0.015)
	Scale - intra-consumer var.	0.10	0.078 (0.029)	0.090 (0.040)	0.072 (0.014)
	Constant – Alt. 1	1.00	1.014 (0.033)	1.021 (0.020)	0.988 (0.011)
	Constant – Alt. 2	0.50	0.522 (0.032)	0.513 (0.020)	0.496 (0.010)
	Constant - Class 1	−0.10	−0.214 (0.265)	0.073 (0.117)	−0.030 (0.070)
	Student - Class 1	1.00	1.193 (0.139)	1.093 (0.087)	0.975 (0.039)
Number of iterations (sampling)			**300,000**	**200,000**	**40,000**

single class logit mixture with inter- and intra-consumer heterogeneity (Becker et al., 2018) (labelled "*Single Class Model*"). We use the baseline sample (5 000 individuals), and the population parameters presented in Tables 1–3.

Table 9 presents the goodness-of-fit statistics for the estimated models on the training data: the log-likelihood, the Akaike Information Criterion (AIC) (Bozdogan, 1987) and the Bayesian Information Criterion (BIC) (Schwarz, 1978). The latter two statistics penalize model complexity as shown in equations (30) and (31):

$$AIC = 2k - 2\ln(\widehat{L}) \tag{30}$$

$$BIC = \ln(N)k - 2\ln(\widehat{L}) \tag{31}$$

Where k is the number of parameters, $\widehat{L}$ is the fitted log-likelihood, and N is the sample size. BIC penalizes the likelihood more aggressively compared to AIC.

As shown in Table 9, the Latent Class Extension achieves a higher log-likelihood and lower AIC and BIC values compared to the Single Class Model, indicating a better fit on the training data.

We can also test the model performance on hold-out data generated for the same individuals (same individual-specific means, but different menus and different menu-specific perturbations). These predictions are conditional on the observed choices; they are calculated using the posterior draws of the individual-specific parameters. Table 10 shows the average predicted probability of the chosen alternative and the error rate in the hold-out data. The error rate is calculated as the percentage of cases where the chosen alternative does not have the highest predicted probability among all three alternatives.

The Latent Class Extension results in a slightly higher predicted probability (by approximately 0.67%) and a lower error rate (by approximately 1.35%). While the magnitude of improvement is relatively small, such improvements can be substantial in some applications (such as personalized price discounts, recommender systems, etc.). In addition, the magnitude of improvement might depend on several factors such as the levels of inter- and intra-consumer heterogeneity, the number of observations per individual, and how distinct the classes are.

Table 8
Monte Carlo analysis.

	Parameter	True value	Single Run	30 Runs
Class 1 Parameters	Transit - population mean	4.00	4.227 (0.116)	3.979 (0.176)
	Bike Sharing - population mean	3.00	3.016 (0.096)	2.999 (0.130)
	On-Demand - population mean	−2.50	−2.578 (0.070)	−2.488 (0.077)
	Transit - inter-consumer var.	1.00	1.020 (0.210)	1.078 (0.318)
	Bike Sharing - inter-consumer var.	1.00	0.755 (0.167)	0.985 (0.184)
	On-Demand - inter-consumer var.	0.50	0.519 (0.073)	0.502 (0.100)
	Transit - intra-consumer var.	0.25	0.289 (0.148)	0.274 (0.125)
	Bike Sharing - intra-consumer var.	0.25	0.317 (0.141)	0.305 (0.101)
	On-Demand - intra-consumer var.	0.25	0.236 (0.062)	0.275 (0.052)
	Transit/Bike Sharing cov. (inter)	0.40	0.532 (0.159)	0.451 (0.235)
	Transit/On-Demand cov. (inter)	−0.4	−0.338 (0.112)	−0.428 (0.160)
	Bike Sharing/On-Demand cov. (inter)	−0.4	−0.360 (0.118)	−0.411 (0.115)
	Transit/Bike Sharing cov. (intra)	0.15	0.134 (0.095)	0.119 (0.083)
	Transit/On-Demand cov. (intra)	−0.15	−0.164 (0.082)	−0.115 (0.071)
	Bike Sharing/On-Demand cov. (intra)	−0.15	−0.077 (0.058)	−0.117 (0.065)
Class 2 Parameters	Transit - population mean	1.00	1.035 (0.075)	1.002 (0.080)
	Bike Sharing - population mean	1.00	1.085 (0.050)	0.996 (0.056)
	On-Demand - population mean	−1.50	−1.526 (0.025)	−1.497 (0.025)
	Transit - inter-consumer var.	0.50	0.668 (0.128)	0.524 (0.095)
	Bike Sharing - inter-consumer var.	0.50	0.693 (0.098)	0.452 (0.087)
	On-Demand - inter-consumer var.	0.25	0.252 (0.025)	0.239 (0.027)
	Transit - intra-consumer var.	0.50	0.430 (0.118)	0.560 (0.134)
	Bike Sharing - intra-consumer var.	0.50	0.356 (0.089)	0.479 (0.144)
	On-Demand - intra-consumer var.	0.50	0.514 (0.034)	0.497 (0.038)
	Transit/Bike Sharing cov. (inter)	0.2	0.341 (0.083)	0.204 (0.087)
	Transit/On-Demand cov. (inter)	−0.2	−0.241 (0.040)	−0.196 (0.038)
	Bike Sharing/On-Demand cov. (inter)	−0.2	−0.248 (0.040)	−0.178 (0.046)
	Transit/Bike Sharing cov. (intra)	0.3	0.183 (0.076)	0.291 (0.097)
	Transit/On-Demand cov. (intra)	−0.2	−0.201 (0.050)	−0.200 (0.054)
	Bike Sharing/On-Demand cov. (intra)	−0.2	−0.250 (0.049)	−0.187 (0.070)
Shared Parameters	Scale - population mean	0.50	0.477 (0.030)	0.507 (0.030)
	Scale - inter-consumer var.	0.25	0.274 (0.038)	0.255 (0.024)
	Scale - intra-consumer var.	0.10	0.090 (0.040)	0.102 (0.034)
	Constant – Alt. 1	1.00	1.021 (0.020)	0.996 (0.023)
	Constant – Alt. 2	0.50	0.513 (0.020)	0.498 (0.022)
	Constant - Class 1	−0.10	0.073 (0.117)	−0.128 (0.194)
	Student - Class 1	1.00	1.093 (0.087)	1.026 (0.093)

Table 9
Goodness-of-fit comparison against the single class model (Becker et al., 2018).

	Latent Class Extension	Single Class Model
Number of parameters	37	20
Log-Likelihood	−22,810.9	−22,988.21
AIC	45,695.8	46,016.4
BIC	46,013.9	46,188.4

Table 10
Predictive power comparison against the single class model.

	Latent Class Extension	Single Class Model
Predicted Probability	74.07%	73.40%
Error Rate	22.36%	23.71%

These results show that when the data consists of multiple classes, each having its own population means and inter- and intra-consumer covariance matrices, assuming normal or log-normal distributions and a single intra-consumer covariance matrix can result in a worse model fit and worse predictions.

In this case, it was not possible to estimate a model with three classes. The latter estimation reduces to two classes as the probability of the third class goes to zero. This can result in the Gibbs sampler breaking down, as the number of individuals assigned to the third class can be zero. This can be prevented by using an informative prior on class membership (e.g. larger concentration parameters α_s), but the estimation still essentially reduces to the two-class model (as the probability of the third class will be extremely small).

5. Real application

5.1. Model estimation

The model described above is applied to the Swiss Route Choice data set (Axhausen et al., 2008), which is publicly available within the R package "Apollo" (Hess and Palma, 2019). The data comes from a stated preferences (SP) survey of public transport route choice conducted in Switzerland, with a sample of 388 respondents faced with 9 choice situations each. Thus, the number of observations is 3 492.

Each choice situation involves two alternatives with four different attributes: travel time (mins.), travel cost in Swiss Francs (CHF), headway (mins.), and the number of transfers. In addition, data on household car availability, income, and trip purpose are available.

The purpose of this application is to analyze the distributions of the willingness-to-pay (WTP) for in-vehicle travel time, waiting time (headway), and the number of transfers, while accounting for inter- and intra-consumer heterogeneity. Car availability and household income are used as predictors of class membership. The utility specifications are shown in equation (32).

$$\begin{aligned} U_{jmn} = \alpha^* \big(-Cost_{jmn} - \exp\left(\beta_{time,mn}\right) \times Time_{jmn} - \exp\left(\beta_{HE,mn}\right) \times HE_{jmn} \\ -\exp\left(\beta_{Transfers}\right) \times Transfers_{jmn} \big) + \varepsilon_{jmn}; \qquad j = 1,2 \end{aligned} \tag{32}$$

Where α is a non-random scale parameter, $Cost_{jmn}$ is the cost of alternative j (CHF), $Time_{jmn}$ and HE_{jmn} are the in-vehicle travel time and headway of alternative j (in hours), and $Transfers_{jmn}$ is the number of transfers of alternative j. The model is specified in the WTP space; the cost coefficient is fixed to -1 and a scale parameter is estimated. Therefore, all the coefficients represent the WTP values for their corresponding attributes.

The class membership model is specified as a binary logit model, with the utility specification shown in equation (33).

$$\begin{aligned} U_1 &= \gamma_0 + \gamma_1 CarAvailability_n + \gamma_2 Income_n + \delta_{n1} \\ U_2 &= 0 + \delta_{n2} \end{aligned} \tag{33}$$

Where $CarAvailability_n$ is a binary indicator of household car availability, $Income_n$ is the income in 10,000's CHF, and δ_{n1} and δ_{n2} are Extreme Value (EV (0,1)) error terms.

This model is estimated with 500,000 Gibbs iterations using R. Two chains are run with different starting values (for calculating the Gelman and Rubin convergence diagnostics), and 100,000 draws are used as burn-in draws. The results are presented in Table 11, showing the posterior means and standard deviations of all the estimated parameters, in addition to the R-hat convergence diagnostics. The R-hat values are all close to 1.0, indicating that the Markov Chains are stationary (the magnitudes of the between-chain and within-chain variances are similar).

Based on the estimates, we can calculate the mean, median, and standard deviation of the WTP values for travel time, headway, and the number of transfers. These are presented in Table 12. These results show that Class 1 (which is more likely to have a car and higher income) has higher WTP values for all three attributes (travel time, waiting time, and number of transfers).

5.2. Model selection

In this section, a similar model is estimated with one class (standard logit mixture) and the AIC and BIC statistics are calculated for both models. The results are presented in Table 13. The log-likelihood of the Latent Class Extension is higher than that of the Single Class Model as expected. The AIC statistic favors the Latent Class Extension, while the BIC statistic (which penalizes model complexity

Table 11
Estimation results using the Swiss Route Choice dataset.

Variable Parameters: WTP Estimates		Class 1		Class 2	
		Posterior mean (Std. dev.)	R-hat	Posterior mean (Std. dev.)	R-hat
Time	Population Mean	3.324 (0.167)	1.09	2.195 (0.200)	1.09
	Inter-consumer variance	0.286 (0.092)	1.01	0.489 (0.205)	1.01
	Intra-consumer variance	0.221 (0.082)	1.00	0.547 (0.275)	1.04
Headway	Population Mean	2.465 (0.280)	1.10	0.554 (0.569)	1.07
	Inter-consumer variance	0.743 (0.340)	1.03	1.314 (0.936)	1.03
	Intra-consumer variance	0.413 (0.226)	1.01	0.570 (0.752)	1.02
Transfers	Population Mean	1.374 (0.181)	1.02	0.579 (0.133)	1.04
	Inter-consumer variance	0.866 (0.376)	1.03	0.284 (0.146)	1.01
	Intra-consumer variance	0.963 (0.339)	1.02	0.240 (0.134)	1.01
Fixed Parameters					
		Posterior mean (Std. dev.)		**R-hat**	
Scale		1.313 (0.162)		1.00	
Constant (class membership)		−0.766 (0.713)		1.08	
Income (class membership)		0.860 (0.460)		1.02	
Car availability (class membership)		0.670 (0.418)		1.09	

Table 12
WTP statistics for the two classes in the Swiss Route dataset.

		Class 1	Class 2
VOT (CHF/hr)	Mean	35.80	15.08
	Median	27.78	8.98
	Standard Deviation	27.98	15.35
VOH (CHF/hr)	Mean	20.97	4.47
	Median	11.76	1.74
	Standard Deviation	22.67	6.95
VOTr (CHF/Transfer)	Mean	9.86	2.32
	Median	3.95	1.78
	Standard Deviation	14.92	1.83

more heavily) favors the Single Class Model.

This indicates that the Latent Class Extension provides a slightly better fit than the Single Class Model. This slight improvement is expected because the estimated distributions are unimodal as shown in Fig. 2, which makes the single class approximation and normality assumptions reasonable in this case. This figure shows the estimated inter-consumer distributions obtained from the Single Class Model and the Latent Class Extension.

6. Discussion

The following sections present a discussion of the applications of the proposed model, its extensions, and limitations.

6.1. Model selection

In estimating such models, considerable care is needed to determine the optimal number of classes. In this paper, we presented two methods: out of sample validation and statistical tests (AIC and BIC), both of which have advantages and disadvantages. AIC and BIC use the training data, but penalize model complexity (i.e., the number of estimated parameters), which limits over-fitting. BIC penalizes more heavily than AIC, which means that these two methods can reach different conclusions (BIC favors more parsimonious models).

On the other hand, out of sample validation is more objective; the selected model should have a better performance on unseen data. However, this requires a separate data set for validation. Depending on the sample size, this might not always be feasible to the modeler; sometimes it might be better to use all the available data in estimation to obtain a better specification and better convergence.

Both of these tests require estimating models with a different number of classes beforehand and deciding on the best model. Another approach is to incorporate a “Chinese Restaurant Process”, in which the number of classes varies at each iteration of the Gibbs sampler. Such models have been proposed by Burda et al. (2008) for logit and probit mixture models (with inter-consumer heterogeneity only).

6.2. Advantages and applications

The proposed model extends the logit double mixture model proposed by Becker et al. (2018) to allow for flexible mixing distributions which can be determined by the data, relaxing the normality assumptions. Our results show that when the data comes from two distinct classes, this model results in a better fit and predictions. Even though the improvement in predictions is small compared to the single class logit mixture model, it can be substantial in some applications. For example, even such small improvements in predictions can drive large revenues in real-time decision support systems, such as recommender systems or personalized incentives (e.g. price discounts). This model can be useful in model application and policy design. Different policies can be tailored to different classes in the population, or even to different individuals (if socio-economic variables are used to predict class membership).

Due to the potentially large number of estimated parameters, the proposed model requires large data sets for estimation, which might not be available in typical stated preferences studies. This model can be applicable to “Big Data” settings such as app-based or web-based systems, where the number of observations is large and data are collected online.

The HB estimator can also be used in online personalization settings in order to estimate and update preferences at the individual level (e.g. in recommender systems). Online estimation of discrete choice models was proposed by Danaf et al. (2019) to update individual preferences after each choice, building on the logit double mixture model in Becker et al. (2018). The online estimator is computationally efficient, because it uses the data of the individual making the choice only in updating his/her individual preferences. This is done by iterating Steps 4 and 5 of the Gibbs sampler described in Section 2.3 (which update the individual- and choice-specific parameters respectively), while assuming that the population parameters (mean and inter- and intra-consumer covariance matrices) are fixed. Periodically, data from multiple individuals are pooled, and the population parameters are updated offline.

The estimator proposed in this paper maintains the online estimation capability; the online Gibbs sampler iterates over steps 4, 6, and 7 described in Section 3.2, which are used to update the individual- and choice-specific parameters as well as the individual-specific class membership indicator after each choice.

Modeling flexible mixing distributions with intra-consumer heterogeneity can also provide insights on model estimation and

Table 13
Goodness-of-fit statistics (AIC and BIC) compared to the single class model in the Swiss Route dataset.

	Latent Class Extension	Single Class Model
Number of Parameters	22	10
Log-likelihood	−1 511.3	−1 537.1
AIC	3 066.5	3 094.3
BIC	3 202.0	3 155.9

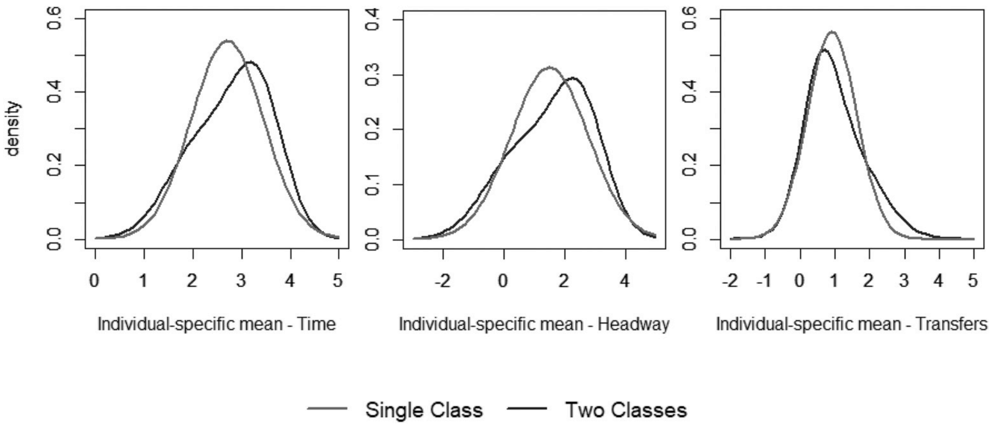

Fig. 2. Distribution of the time, headway, and transfers parameters obtained with one class and two classes in the Swiss Route dataset.

response biases in SP surveys. For example, high magnitudes of inter- or intra-consumer heterogeneity in some attributes might indicate omitted variables causing this heterogeneity.

According to Ben-Akiva et al. (2019), latent class models can be used to correct for several SP biases such as random responses, attribute non-attendance, and different choice protocols. While the proposed model uses latent classes to represent flexible mixing distributions, it can be extended to account for other SP biases. For example, the systematic utility equations of one of the classes can be specified as zeroes (to correct for random response bias) or constants only, or the distributions of some parameters can include point masses at zero (to represent attribute non-attendance).

6.3. Convergence

The main drawback of the proposed model is that it is computationally burdensome to estimate. In the Monte Carlo and SP applications in Sections 4 and 5, estimation takes twice the time required for estimating the single class logit mixture model with inter- and intra-consumer heterogeneity.

In addition, estimation using Gibbs sampling might get stuck in some regions of the posterior distribution (corresponding to local optima). For example, if one of the classes has a very tight distribution of inter- or intra-consumer heterogeneity, the probability of individuals switching to that class (in Step 5 of the Gibbs sampler) will be very low.

These models should be estimated using different starting values, and multiple chains should be run in order to validate that they converge to the same posterior distributions (e.g. using the Gelman and Rubin (1992) diagnostic). Whenever available, good informative priors (or weakly informative priors) can be used to achieve better and faster convergence. For example, good priors on the covariance matrices can guarantee that variances do not approach zero, in order to avoid the problem described above (where the Gibbs sampler is stuck because of small variances).

Finally, estimation can be susceptible to label switching, especially with small sample sizes, or when the mixture components are not well separated (in the latter case, the latent class extension might not be needed, as a single class might be sufficient to represent heterogeneity). Label switching was encountered in our application to SP data, particularly when class membership was represented using a Multinomial-Dirichlet mixture model. However, this was mitigated by modeling class membership as a function of socio-demographic characteristics. With larger sample sizes (i.e., in the synthetic data), label switching was only observed before stationarity. Therefore, the class labels can switch across different estimations depending on the starting values, but this will not affect the overall estimation results (as it happens between different estimations, and not within an estimation).

7. Conclusion

This paper presented an extension to the logit mixture model with inter- and intra-consumer heterogeneity, which overcomes the normality assumptions imposed on the mixing distributions, and allows different individuals to have different intra-consumer covariance matrices. According to Vij and Krueger (2017), semi-parametric mixture models allow for modeling complex patterns of heterogeneity that cannot be captured using uni-modal distributions.

We also presented an HB estimator consisting of seven Gibbs layers with an embedded MH algorithm. This estimator extends the HB estimator proposed by Becker et al. (2018) and Ben-Akiva et al. (2019) by adding a latent class model, either with fixed probabilities or using a parametrized class membership model. The proposed model and the HB estimator were validated using real and synthetic data sets, and the models were evaluated using the AIC and BIC statistics and out-of-sample validation.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript

References

Allenby, G.M., Rossi, P.E., 1998. Marketing models of consumer heterogeneity. Journal of econometrics 89 (1–2), 57–78.
Andrews, R.L., Ainslie, A., Currim, I.S., 2002. An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. J. Mark. Res. 39 (4), 479–487.
Akinc, D., Vandebroek, M., 2018. Bayesian estimation of mixed logit models: selecting an appropriate prior for the covariance matrix. Journal of choice modelling 29, 133–151.
Allenby, G., 1997. An introduction to hierarchical Bayesian modeling. In: Tutorial notes, Advanced Research Techniques. Forum, American Marketing Association.
Allenby, G.M., Rossi, P.E., 1998. Marketing models of consumer heterogeneity. J. Econom. 89 (1–2), 57–78.
Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J.J., Bierlaire, M., 2008. State of the art estimates of the swiss value of travel time savings. Transp. Policy 15 (3), 173–185.
Bajari, P., Fox, J.T., Ryan, S.P., 2007. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. Am. Econ. Rev. 97 (2), 459–463.
Becker, F., Danaf, M., Song, X., Atasoy, B., Ben-Akiva, M., 2018. Hierarchical Bayes estimator of a logit mixture with inter- and intra-consumer heterogeneity. Transp. Res. Part B Methodol. 117 (A), 1–17.
Ben-Akiva, M., McFadden, D., Train, K., 2019. Foundations of stated preference elicitation: consumer behavior and choice-based conjoint analysis. Foundations and Trends in Econometrics 10 (1–2), 1–144. https://doi.org/10.1561/0800000036.
Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transp. Res. Part B Methodol. 45 (7), 923–939.
Bhat, C.R., Castelar, S., 2002. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. Transp. Res. Part B Methodol. 36 (7), 593–616.
Bhat, C.R., Sardesai, R., 2006. The impact of stop-making and travel time reliability on commute mode choice. Transp. Res. Part B Methodol. 40 (9), 709–730.
Bhat, C.R., Sidharthan, R., 2011. A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. Transp. Res. Part B Methodol. 45 (7), 940–953.
Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52 (3), 345–370.
Bujosa, A., Riera, A., Hicks, R.L., 2010. Combining discrete and continuous representations of preference heterogeneity: a latent class approach. Environ. Resour. Econ. 47 (4), 477–493.
Burda, M., Harding, M., Hausman, J., 2008. A Bayesian mixed logit–probit model for multinomial choice. J. Econom. 147 (2), 232–246.
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.
Elrod, T., Keane, M.P., 1995. A factor-analytic probit model for representing the market structure in panel data. J. Mark. Res. 1–16.
Fosgerau, M., Bierlaire, M., 2007. A practical test for the choice of mixing distribution in discrete choice models. Transp. Res. Part B Methodol. 41 (7), 784–794.
Fox, J.T., Kim, K.I., Ryan, S.P., Bajari, P., 2011. A simple estimator for the distribution of random coefficients. Quantitative Economics 2 (3), 381–418.
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7 (4), 457–472.
Geweke, J., Keane, M., 2001. Computationally intensive methods for integration in econometrics. In: Handbook of Econometrics, vol. 5. Elsevier, pp. 3463–3568.
Geweke, J., Keane, M., 2007. Smoothly mixing regressions. J. Econom. 138 (1), 252–290.
Greene, W.H., Hensher, D.A., 2013. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. Appl. Econ. 45 (14), 1897–1902.
Hess, S., 2014. Latent class structures: taste heterogeneity and beyond. In: Handbook of Choice Modelling. Edward Elgar Publishing Cheltenham, pp. 311–329 (Chapter 14).
Hess, S., Palma, D., 2019. Apollo: a Flexible, Powerful and Customisable Freeware Package for Choice Model Estimation and Application.
Hess, S., Rose, J.M., 2009. Allowing for intra-respondent variations in coefficients estimated on repeated choice data. Transportation Research Part B: Methodological 43 (6), 708–719.
Hess, S., Train, K., 2011. Recovery of inter- and intra-consumer heterogeneity using mixed logit models. Transportation Research Part B: Methodological 45 (7), 973–990.
Huang, A., Wand, M.P., 2013. Simple marginally noninformative prior distributions for covariance matrices. Bayesian Analysis 8 (2), 439–452.
Keane, M., Wasi, N., 2013. Comparing alternative models of heterogeneity in consumer choice behavior. J. Appl. Econom. 28 (6), 1018–1045.
Krueger, R., Vij, A., Rashidi, T.H., 2018. A Dirichlet Process Mixture Model of Discrete Choice arXiv preprint arXiv:1801.06296.
Rossi, P., Allenby, G., McCulloch, R., 2005. Bayesian Statistics and Marketing. John Wiley and Sons, Hoboken, N.J.
Sarrias, M., Daziano, R.A., 2017. Multinomial logit models with continuous and discrete individual heterogeneity in R: the gmnl package. J. Stat. Softw. 79 (2), 1–46.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Train, K., 2008. EM algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling 1 (1), 40–69.
Train, K.E., 2009. Discrete choice methods with simulation. Cambridge university press.
Train, K., 2016. Mixed logit with a flexible mixing distribution. Journal of choice modelling 19, 40–53.
Vij, A., Krueger, R., 2017. Random taste heterogeneity in discrete choice models: flexible nonparametric finite mixture distributions. Transp. Res. Part B Methodol. 106, 76–101.
Yáñez, M., Cherchi, E., Heydecker, B., Dios-Ortuzar, J., 2011. On the treatment of repeated observations in panel data: efficiency of mixed logit parameter estimates. Networks and Spatial Economics 11 (3), 393–418.