Evaluating the predictive abilities of mixed logit models with unobserved inter- and intra-individual heterogeneity

Rico Krueger [*] Michel Bierlaire [*] Ricardo A. Daziano [†]
Taha H. Rashidi [‡] Prateek Bansal [§]

June 17, 2021

Report TRANSP-OR 210617
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*]École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, {rico.krueger,michel.bierlaire}@epfl.ch

[†]School of Civil and Environmental Engineering, Cornell University, United States, {daziano@cornell.edu}

[‡]Research Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering, University of New South Wales, Australia, {rashidi@unsw.edu.au}

[§]Transport Strategy Centre, Department of Civil and Environmental Engineering, Imperial College London, UK, {prateek.bansal@imperial.ac.uk}

Abstract

Mixed logit models with unobserved inter- and intra-individual heterogeneity hierarchically extend standard mixed logit models by allowing tastes to vary randomly both across individuals and across choice situations encountered by the same individual. Recent work advocates using these models in choice-based recommender systems under the premise that mixed logit models with unobserved inter- and intra-individual heterogeneity afford personalised preference estimation and prediction. In this study, we evaluate the ability of mixed logit with unobserved inter- and intra-individual heterogeneity to produce accurate individual-level predictions of choice behaviour. Using simulated and real data, we show that mixed logit models with unobserved inter- and intra-individual heterogeneity do not provide significant improvements in choice prediction accuracy over standard mixed logit models, which only account for inter-individual taste variation. We make these observations even in scenarios with high levels of intra-individual taste variation and when the number of choice situations per decision-maker is large. Also, the estimation of mixed logit with unobserved inter- and intra-individual heterogeneity requires at least seven times as much computation time as the estimation of standard mixed logit. Drawing from recent advances in machine learning and econometrics, we discuss alternative modelling approaches that can capture richer dependencies between decision-makers, alternatives and attributes.

Keywords: mixed logit; unobserved heterogeneity; recommender systems.

1 Introduction

Mixed random utility models such as mixed logit (McFadden and Train, 2000) provide a powerful framework to account for unobserved taste heterogeneity in discrete choice models. When longitudinal choice data are analysed using mixed random utility models, it is standard practice to assume that tastes vary randomly across decision-makers but not across choice situations encountered by the same individual (Revelt and Train, 1998). The implicit assumption underlying this treatment of unobserved heterogeneity is that an individual's tastes are unique and stable (Stigler and Becker, 1977). However, contrasting views of preference formation postulate that preferences are constructed in an ad-hoc manner at the moment of choice (Bettman et al., 1998) or learnt and discovered through experience (Kivetz et al., 2008).

From a behavioural perspective, these alternative views of preference formation justify accounting for both inter- and intra-individual random heterogeneity in discrete choice models (also see Hess and Giergiczny, 2015). A straightforward way to accommodate unobserved inter- and intra-individual heterogeneity in mixed random utility models is to augment a normal mixing distribution in a hierarchical fashion such that case-specific taste parameters are generated as normal perturbations around individual-specific taste parameters (see Becker et al., 2018, Ben-Akiva et al., 2019, Bhat and Castelar, 2002, Bhat and Sardesai, 2006, Bhat and Sidharthan, 2011, Danaf et al., 2019, Hess and Giergiczny, 2015, Hess and Rose, 2009, Hess and Train, 2011, Xie et al., 2020, Yáñez et al., 2011).

Originally, mixed logit models with unobserved inter- and intra-individual heterogeneity were primarily used as variance decomposition techniques in order to separate unobserved taste variation into inter- and intra-individual terms. Yet, recent work advocates using these methods in choice-based recommender systems under the premise that mixed logit models with unobserved inter- and intra-individual heterogeneity afford personalised preference estimation and prediction (Danaf et al., 2019, Xie et al., 2020). These studies demonstrate that mixed logit models with unobserved inter- and intra-individual heterogeneity outperform standard logit models at out-of-sample prediction, both unconditionally (i.e. non-personalised inter-individual prediction for respondents without a history of past choices) and conditionally (i.e. personalised intra-individual prediction for respondents with a history of past choices) individuals. However, these studies do not draw comparisons with standard mixed logit models, which only account for inter-individual heterogeneity. Danaf et al. (2019) and Xie et al. (2020) contrast non-personalised (unconditional) and personalised (conditional) predicted choice probabilities of mixed logit with inter- and intra-individual heterogeneity. As expected, they conclude that personalisation improves the conditional prediction accuracy. However, unconditional choice probabilities of mixed logit with inter- and intra-individual heterogeneity are not the same as conditional choice

probabilities of standard mixed logit models.

In this research note, we evaluate the ability of mixed logit models with unobserved inter- and intra-individual heterogeneity to provide personalised predictions of choice behaviour. Using simulated and real data, we show that mixed logit models with unobserved inter- and intra-individual heterogeneity provide only marginal gains in terms of conditional predictions over simpler, computationally less expensive mixed logit models with only inter-individual heterogeneity. In light of these findings and informed by recent advances at the intersection of machine learning and econometrics, we then discuss alternative approaches adopted in recommender systems to generate personalised predictions with random utility models.

With the growing availability of dynamic panel data sets, recommender systems are increasingly employed to increase user satisfaction and lower search costs by helping users to navigate complex goods and service systems such as Internet marketplaces and smart mobility (Ansari et al., 2000, Lu et al., 2015, Song et al., 2018). An example of a recommender system is a mobile local search-and-discovery application that provides users with personalised recommendations of places like restaurants based on user characteristics and previous visits (Kim, 2015). Accurate methods for personalised preference estimation and prediction lie at the heart of successful recommender systems (Ansari et al., 2000). Unlike standard recommendation methods such as collaborative and content-based filtering, discrete choice models can be employed even when the choice set is not persistent (Danaf et al., 2019). Consequently, there is a synergy between the methods adopted in recommender systems and discrete choice models, because the success of both approaches depends critically on the ability to capture rich dependencies between individuals, alternatives and attributes (Jiang et al., 2014).

Several remarks about the focus of our contribution are in order: First, we emphasise that the main focus of our contribution is on evaluating the predictive abilities of mixed logit with unobserved inter- and intra-individual heterogeneity and not on understanding behaviour. At the same time, our analysis includes comparisons of maximum simulated likelihood and Bayes estimators for mixed logit models with unobserved inter- and intra-individual heterogeneity. These comparisons are also relevant to researchers who are mainly interested in using the model to explain behaviour. Furthermore, we discuss several emerging approaches from the recommender systems literature which can be incorporated into the random utility maximisation framework in order to improve the conditional prediction accuracy of discrete choice models. This research direction is timely and relevant because the methods adopted in recommender systems offer flexible, parametric representations of the dependencies between individuals, alternatives and attributes. Until recently, these innovative models were expensive to estimate using standard methods due to their large parameter spaces. However, emerging approximate inference methods such as variational inference offer a drastic reduction of the

computational burden associated with the estimation of such complex probabilistic models (Bansal et al., 2020, Hosseini et al., 2018).
We organise the remainder of this research note as follows. First, we introduce mixed logit with unobserved inter- and intra-individual heterogeneity (Section 2). Next, we present the simulation study and the real data application (Sections 3 and 4). Then, we provide an extended discussion of alternative modelling approaches (Section 5), and finally, we conclude (Section 6).

2 Methodology

2.1 Model formulation

Mixed logit with unobserved inter- and intra-individual heterogeneity (in particular Hess and Rose, 2009, Hess and Train, 2011) is established as follows: In choice situation $t \in \{1, \dots T\}$, a decision-maker $n \in \{1, \dots N\}$ derives utility

$$U_{ntj} = V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt}) + \varepsilon_{ntj} \tag{1}$$

from alternative j in the set $\mathcal{C} = \{1, \dots, J\}$. Here, $V()$ denotes the deterministic aspect of utility, $\mathbf{X}_{ntj}$ is a vector of covariates, $\boldsymbol{\beta}_{nt}$ is a collection of taste parameters, and ε_{ntj} is a stochastic disturbance. We obtain the logit model under the assumption that ε_{ntj} is independently and identically distributed according to Gumbel$(0, 1)$ across decision-makers n, choice situations t and alternatives j. Consequently, the probability that decision-maker n chooses alternative $j \in \mathcal{C}$ in choice situation t can be expressed as

$$P(y_{nt} = j|\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt},) = \frac{e^{V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt})}}{\sum_{j' \in \mathcal{C}} e^{V(\mathbf{X}_{ntj'}, \boldsymbol{\beta}_{nt})}}, \tag{2}$$

where the random variable $y_{nt} \in \mathcal{C}$ indicates the chosen alternative.
The distinguishing feature of mixed logit with unobserved inter- and intra-individual heterogeneity is that the taste parameters $\boldsymbol{\beta}_{nt}$ are case-specific. More specifically, $\boldsymbol{\beta}_{nt}$ is a normal perturbation around an individual-specific parameter $\boldsymbol{\mu}_n$, i.e. $\boldsymbol{\beta}_{nt} \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)$ for $t = 1, \dots, T$, where $\boldsymbol{\Sigma}_W$ is a full covariance matrix. The distribution of the individual-specific parameter $\boldsymbol{\mu}_n$ is then also multivariate normal, i.e. $\boldsymbol{\mu}_n \sim N(\zeta, \boldsymbol{\Sigma}_B)$ for $n = 1, \dots, N$, where ζ is a mean vector and $\boldsymbol{\Sigma}_B$ is a full covariance matrix. In contradistinction, the standard panel estimator for mixed logit assumes taste homogeneity across replications, i.e. $\boldsymbol{\beta}_{nt} = \boldsymbol{\beta}_n \ \forall t \in \{1, \dots, T\}$, in order to capture inter-individual taste heterogeneity and to allow for dependence across repeated observations (Revelt and Train, 1998).
The generally adopted labels inter- and intra-individual heterogeneity may falsely suggest that inferences are performed at the individual level. However, this is not the case. Compared to standard mixed logit, mixed logit with inter- and intra-individual heterogeneity has one more level to capture taste variation across

choice situations. We learn about taste variation at the inter-individual level using information about differences across respondents in a longitudinal dataset. However, since $\boldsymbol{\Sigma}_W$ is generic, we learn about taste variation at the intra-individual level using information about differences across all choices from all respondents rather than across choices from only one respondent.

2.2 Estimation

Mixed logit with unobserved inter- and intra-individual heterogeneity can be estimated using either classical maximum simulated likelihood (MSL) or Bayesian Markov chain Monte Carlo (MCMC) methods. In what follows, we describe both estimation approaches.

2.2.1 Maximum simulated likelihood (MSL)

In MSL estimation, the parameters $\theta = \{\zeta, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$ are treated as fixed, unknown quantities. Point estimates of θ are obtained via maximisation of the unconditional log-likelihood, whereby the optimisation is in fact performed with respect to the Cholesky factors $\{\mathbf{L}_B, \mathbf{L}_W\}$ of $\{\boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$ in order to maintain positive-definiteness of the covariance matrices. Unlike in Bayesian estimation, the stochastic parameters $\boldsymbol{\mu}_n$ and $\boldsymbol{\beta}_{nt}$ are not directly estimated, because they are integrated out in the simulation of the unconditional log-likelihood.
To formulate the unconditional log-likelihood, we define $\boldsymbol{\beta}_{nt} = \boldsymbol{\mu}_n + \boldsymbol{\gamma}_{nt}$, where $\boldsymbol{\mu}_n \sim N(\zeta, \boldsymbol{\Sigma}_B)$ is an individual-specific random parameter with density $f(\boldsymbol{\mu}_n|\zeta, \boldsymbol{\Sigma}_B)$, and where $\boldsymbol{\gamma}_{nt} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_W)$ is a case-specific random parameter with density $f(\boldsymbol{\gamma}_{nt}|\boldsymbol{\Sigma}_W)$. We then obtain the unconditional log-likelihood by marginalising out the stochastic parameters $\boldsymbol{\mu}_n$ and $\boldsymbol{\beta}_{nt}$. We have

$$LL(\theta) = \sum_{n=1}^{N} \ln \left(\int \prod_{t=1}^{T} \left(\int P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt}) f(\boldsymbol{\gamma}_{nt}|\boldsymbol{\Sigma}_W) d\boldsymbol{\gamma}_{nt} \right) f(\boldsymbol{\mu}_n|\zeta, \boldsymbol{\Sigma}_B) d\boldsymbol{\mu}_n \right). \quad (3)$$

Since the integrals in (3) are not analytically tractable, we resort to simulation to approximate the log-likelihood. The simulated log-likelihood is given by

$$SLL(\theta) = \sum_{n=1}^{N} \ln \left(\frac{1}{D} \sum_{d=1}^{D} \prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \right) \right), \quad (4)$$

where we define $\boldsymbol{\beta}_{nt,dr} = \zeta + \mathbf{L}_B \xi_{n,d} + \mathbf{L}_W \xi_{nt,r}$. Here, $\xi_{n,d}$ and $\xi_{nt,r}$ denote standard normal simulation draws. For each decision-maker, we take D draws to marginalise out $\boldsymbol{\mu}_n$ and R draws to marginalise out $\boldsymbol{\gamma}_{nt}$. A point estimate $\widehat{\theta}$ is then given by

$$\widehat{\theta} = \arg\max_{\theta} SLL(\theta). \quad (5)$$

The optimisation problem defined in (5) can be solved using quasi-Newton methods, which exploit the gradient of the objective function to find a local optimum.

Numerical gradient approximations are computationally expensive, as they incur many evaluations of the objective function. However, computation times of quasi-Newton methods can be drastically reduced, if analytical gradients of the objective are provided. In the case of mixed logit with unobserved inter- and intra-individual heterogeneity, the two levels of integration in the approximation of the unconditional log-likelihood impose a substantial computational burden. Consequently, efficient optimisation routines are critical for moderating estimation times. In Appendix A.1, we present the analytical gradient of (4).

After computing the point estimate $\widehat{\theta} = \{\widehat{\zeta}, \widehat{\Sigma}_B, \widehat{\Sigma}_W\}$, we can obtain the posterior distribution of μ_n using Bayes theorem (Revelt and Train, 2000, Train, 2009). We have

$$P(\mu_n|y_n, X_n, \widehat{\theta}) = \frac{P(y_n|X_n, \mu_n, \widehat{\Sigma}_W) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B)}{\int P(y_n|X_n, \mu_n, \widehat{\Sigma}_W) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B) d\mu_n}. \tag{6}$$

The mean of this posterior distribution is given by $\check{\mu}_n = \int \mu_n P(\mu_n|y_n, X_n, \widehat{\theta}) d\mu_n$. Consequently, we have

$$\begin{aligned} \check{\mu}_n &= \frac{\int \mu_n P(y_n|X_n, \mu_n, \widehat{\Sigma}_W) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B) d\mu_n}{\int P(y_n|X_n, \mu_n, \widehat{\Sigma}_W) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B) d\mu_n} \\ &= \frac{\int \mu_n \left(\prod_{t=1}^{T} \int P(y_{nt}|X_{nt}, \beta_{nt}) f(\gamma_{nt}|\Sigma_W) d\gamma_{nt} \right) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B) d\mu_n}{\int \left(\prod_{t=1}^{T} \int P(y_{nt}|X_{nt}, \beta_{nt}) f(\gamma_{nt}|\Sigma_W) d\gamma_{nt} \right) f(\mu_n|\widehat{\zeta}, \widehat{\Sigma}_B) d\mu_n}. \end{aligned} \tag{7}$$

Since the integrals in (7) are not analytically tractable, we resort to simulation to approximate the posterior mean. The simulated posterior mean $\widehat{\mu}_n$ is given by

$$\widehat{\mu}_n = \sum_{d=1}^{D} w_d \mu_{n,d} \tag{8}$$

with

$$w_d = \frac{\prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|X_{nt}, \beta_{nt,dr}) \right)}{\sum_{d'=1}^{D} \prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|X_{nt}, \beta_{nt,d'r}) \right)}. \tag{9}$$

2.2.2 Markov chain Monte Carlo (MCMC)

The goal of Bayesian estimation is to infer the posterior distribution of all model parameters $\{\zeta, \Sigma_B, \Sigma_W, \mu, \beta\}$. Thus, unlike in MSL estimation, posterior samples of μ_n and β_{nt} are directly obtained along with posterior samples of the other parameters.

The Bayesian approach entails the specification of full probability model for all parameters. Therefore, we also need to assign priors to $\{\zeta, \Sigma_B, \Sigma_W\}$. We use a vague normal prior for ζ, i.e. $\zeta \sim N(\lambda_0, \Lambda_0)$, a half-t prior for Σ_B, Σ_W. The latter is selected because of its superior non-informativity properties compared

to alternative prior specifications for covariance matrices (Akinc and Vandebroek, 2018, Huang and Wand, 2013). The half-t prior is defined hierarchically: It consists of an inverse Wishart prior for $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} \sim \text{IW}(\nu + K - 1, 2\nu\boldsymbol{\Delta})$, where ν is a known hyper-parameter and K denotes the number of random parameters. $\boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_K)$ is a diagonal matrix with elements δ_k distributed Gamma $\left(\frac{1}{2}, \frac{1}{a_k^2}\right)$. Stated succinctly, the full generative process of mixed logit with unobserved inter- and intra-individual heterogeneity is as follows:

$$\delta_{B,k}|a_{B,k} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{a_{B,k}^2}\right), k = 1, \ldots, K, \tag{10}$$

$$\delta_{W,k}|a_{W,k} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{a_{W,k}^2}\right), k = 1, \ldots, K, \tag{11}$$

$$\boldsymbol{\Sigma}_B|\nu_B, \boldsymbol{\delta}_B \sim \text{IW}\left(\nu_B + K - 1, 2\nu_B \text{diag}(\boldsymbol{\delta}_B)\right), \quad \boldsymbol{\delta}_B = \begin{bmatrix}\delta_{B,1} & \ldots & \delta_{B,K}\end{bmatrix}^\top \tag{12}$$

$$\boldsymbol{\Sigma}_W|\nu_W, \boldsymbol{\delta}_W \sim \text{IW}\left(\nu_W + K - 1, 2\nu_W \text{diag}(\boldsymbol{\delta}_W)\right), \quad \boldsymbol{\delta}_W = \begin{bmatrix}\delta_{W,1} & \ldots & \delta_{W,K}\end{bmatrix}^\top \tag{13}$$

$$\boldsymbol{\zeta}|\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0 \sim N(\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0) \tag{14}$$

$$\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B \sim N(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B), n = 1, \ldots, N, \tag{15}$$

$$\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W), n = 1, \ldots, N, t = 1, \ldots, T, \tag{16}$$

$$y_{nt}|\boldsymbol{\beta}_{nt}, \mathbf{X}_{nt} \sim \text{Logit}\left(V(\boldsymbol{\beta}_{nt}, \mathbf{X}_{nt})\right), n = 1, \ldots, N, \ t = 1, \ldots, T, \tag{17}$$

where $\{\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0, \nu_B, \nu_W, \mathbf{a}_B, \mathbf{a}_W\}$ are known hyper-parameters, and $\theta = \{\boldsymbol{\delta}_B, \boldsymbol{\delta}_W, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W, \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\beta}\}$ are the model parameters whose posterior distribution we wish to estimate. The generative process given in (10)–(17) implies the following joint distribution:

$$\begin{aligned} P(\mathbf{y}, \theta) = &\left(\prod_{n=1}^{N}\prod_{t=1}^{T} P(y_{nt}|\boldsymbol{\beta}_{nt}, \mathbf{X}_{nt}) P(\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)\right)\left(\prod_{n=1}^{N} P(\boldsymbol{\mu}_n|\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B)\right) \\ &P(\boldsymbol{\zeta}|\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_0) P(\boldsymbol{\Sigma}_B|\omega_B, \mathbf{B}_B)\left(\prod_{k=1}^{K} P(\delta_{B,k}|s, r_{B,k})\right) \\ &P(\boldsymbol{\Sigma}_W|\omega_W, \mathbf{B}_W)\left(\prod_{k=1}^{K} P(\delta_{W,k}|s, r_{W,k})\right) \end{aligned} \tag{18}$$

where $\omega_B = \nu_B + K - 1$, $\mathbf{B}_B = 2\nu_B \text{diag}(\boldsymbol{\delta}_B)$, $\omega_W = \nu_W + K - 1$, $\mathbf{B}_W = 2\nu_W \text{diag}(\boldsymbol{\delta}_W)$, $s = \frac{1}{2}$, $r_{B,k} = a_{B,k}^{-2}$ and $r_{W,k} = a_{W,k}^{-2}$. By Bayes' rule, the posterior distribution of interest is given by

$$P(\theta|\mathbf{y}) = \frac{P(\mathbf{y}, \theta)}{\int P(\mathbf{y}, \theta) d\theta} \propto P(\mathbf{y}, \theta). \tag{19}$$

Exact inference of this posterior distribution is not possible, because the model evidence $\int P(\mathbf{y}, \theta) d\theta$ is not analytically tractable. Hence, we resort to approximate inference.

The central idea of MCMC is to approximate a posterior distribution through samples from a Markov chain whose stationary distribution is the target distribution. Gibbs sampling constructs such a Markov chain by iteratively sampling from the conditional posterior distributions of blocks of model parameters. Becker et al. (2018) devise a Gibbs sampler for posterior inference in mixed logit with unobserved inter- and intra-individual heterogeneity. In Appendix A.2, we present one iteration of the sampler. A key feature of the algorithm is that posterior samples of $\boldsymbol{\mu}_n$ are directly obtained.

3 Simulation study

In this section, we present an extensive simulation evaluation of mixed logit with unobserved inter- and intra-individual heterogeneity. The model is benchmarked against simpler conditional and mixed logit models in terms of estimation time, estimation accuracy and out-of-sample predictive accuracy. In addition, we compare the performance of the MSL and MCMC estimators for mixed logit with unobserved inter- and intra-individual heterogeneity.

3.1 Data and experimental setup

For the simulation study, we rely on synthetic choice data, which we generate as follows: The choice sets comprise three unlabelled alternatives, which are characterised by four attributes. Decision-makers are assumed to be utility maximisers and to evaluate the alternatives based on the utility specification

$$U_{ntj} = \mathbf{X}_{ntj}^{\top}\boldsymbol{\beta}_{nt} + \varepsilon_{ntj}. \tag{20}$$

For the generation of the taste parameters $\boldsymbol{\beta}_{nt}$, we consider two scenarios, in which the proportion of the total variance that is due to intra-individual taste heterogeneity is varied. In the two scenarios, $\boldsymbol{\beta}_{nt}$ is drawn via the following process:

$$\boldsymbol{\mu}_n|\zeta, \boldsymbol{\Sigma}_B \sim N(\zeta, \boldsymbol{\Sigma}_B), n = 1, \ldots, N, \tag{21}$$

$$\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W), n = 1, \ldots, N, t = 1, \ldots, T, \tag{22}$$

where $\boldsymbol{\Sigma}_B = \text{diag}(\boldsymbol{\sigma}_B)\boldsymbol{\Omega}_B\text{diag}(\boldsymbol{\sigma}_B)$ and $\boldsymbol{\Sigma}_W = \text{diag}(\boldsymbol{\sigma}_W)\boldsymbol{\Omega}_W\text{diag}(\boldsymbol{\sigma}_W)$. Here, $\{\boldsymbol{\sigma}_B, \boldsymbol{\sigma}_W\}$ represent standard deviation vectors and $\{\boldsymbol{\Omega}_B, \boldsymbol{\Omega}_W\}$ are correlation matrices. The assumed values of ζ, $\boldsymbol{\Omega}_B$ and $\boldsymbol{\Omega}_W$ are enumerated in Appendix B. We define $\sigma_B^2 = 2 \cdot (1 - \alpha) \cdot |\zeta|$ and $\sigma_W^2 = 2 \cdot \alpha \cdot |\zeta|$ with $\alpha \in [0, 1]$, i.e. the total variance of each random parameter is twice the absolute value of its mean, and a proportion α of the total variance is due to intra-individual taste variation. In scenario 1, we set $\alpha = 0.3$, and in scenario 2, we set $\alpha = 0.7$. In both scenarios, the alternative-specific attributes $\mathbf{X}_{ntj}$ are drawn from Uniform$(0, 2)$, which implies an error rate of approximately 20%, i.e. in one fifth of the cases decision-makers deviate from

the systematically best alternative due to the stochastic utility component. In both scenarios, we further set $N = 1000$ and let T take a value in $\{10, 20\}$. For each experimental scenario and for each value of T, we consider 20 replications, whereby the data for each replication are generated using a different random seed.

3.2 Accuracy assessment

We evaluate the accuracy of the estimation approaches in terms of their ability to recover parameters in finite samples and their out-of-sample predictive accuracy.

3.2.1 Parameter recovery

To assess how well the estimation approaches perform at recovering parameters, we calculate the root mean square error (RMSE) for selected parameters, namely the mean vector ζ and the unique elements $\{\Sigma_{B,U}, \Sigma_{W,U}\}$ of the covariance matrices $\{\Sigma_B, \Sigma_W\}$. Given a collection of parameters θ and its estimate $\widehat{\theta}$, RMSE is defined as $\text{RMSE}(\theta) = \sqrt{\frac{1}{M}(\widehat{\theta} - \theta)^\top(\widehat{\theta} - \theta)}$, where M denotes the total number of scalar parameters collected in θ. For MSL, point estimates of ζ, Σ_B and Σ_W are directly obtained. For MCMC, estimates of the parameters of interest are given by the means of the respective posterior draws. As our aim is to evaluate how well the estimation methods perform at recovering the distributions of the realised individual- and observation-specific parameters $\{\mu, \beta\}$, we use the sample mean $\bar{\zeta} = \frac{1}{N}\sum_{n=1}^{N}\mu_n$ and the sample covariances $\bar{\Sigma}_B = \frac{1}{N}\sum_{n=1}^{N}(\mu_n - \bar{\zeta})(\mu_n - \bar{\zeta})^\top$ and $\bar{\Sigma}_W = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(\beta_{nt} - \mu_n)(\beta_{nt} - \mu_n)^\top$ as true parameter values for ζ, Σ_B and Σ_W

3.2.2 Predictive accuracy

We consider two out-of-sample prediction scenarios. In the first scenario, we predict choice probabilities for a new set of individuals without a history of past choices, i.e. we predict *unconditionally* on an individual's past choices. To that end, we generate a test set consisting of 100 observations from 100 new individuals along with each training sample. The realised choices and attributes of this sample are denoted by y^*_{nt} and X^*_{nt}. In the second scenario, we predict choice probabilities for new choice sets for individuals who are already in the training sample and thus have a record of past choices, i.e. we predict *conditionally* on an individual's past choices. To that end, we create another test set by generating additional choice sets for 100 individuals from the training sample. The realised choices and attributes of this sample are denoted by $y^\dagger_{nt}$ and $X^\dagger_{nt}$. We let $T = 1$ in both validation samples.

For mixed logit with unobserved inter- and intra-individual heterogeneity, the estimated predicted choice probabilities for the *unconditional* prediction scenario

are given by

$$\widehat{P}(y^*_{nt}|\mathbf{X}^*_{nt}, \mathbf{y}) = \int \left(\int P(y^*_{nt}|\mathbf{X}^*_{nt}, \boldsymbol{\beta}_{nt}) f(\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \widehat{\boldsymbol{\Sigma}}_W) d\boldsymbol{\beta}_{nt} \right) f(\boldsymbol{\mu}_n|\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\Sigma}}_B) d\boldsymbol{\mu}_n, \quad (23)$$

where $\widehat{\boldsymbol{\zeta}}$, $\widehat{\boldsymbol{\Sigma}}_B$ and $\widehat{\boldsymbol{\Sigma}}_W$ denote the posterior means of $\boldsymbol{\zeta}$, $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$, respectively. The estimated predicted choice probabilities for the *conditional* prediction scenario are given by

$$\widehat{P}(y^\dagger_{nt}|\mathbf{X}^\dagger_{nt}, \mathbf{y}) = \int P(y^\dagger_{nt}|\mathbf{X}^\dagger_{nt}, \boldsymbol{\beta}_{nt}) f(\boldsymbol{\beta}_{nt}|\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_W) d\boldsymbol{\beta}_{nt}, \quad (24)$$

where $\widehat{\boldsymbol{\mu}}_n$ and $\widehat{\boldsymbol{\Sigma}}_W$ denote the posterior means of $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_W$, respectively. Recall that in MCMC, the posterior distribution of $\boldsymbol{\mu}_n$ is directly estimated, whereas for MSL, we obtain $\widehat{\boldsymbol{\mu}}_n$ using (8). Expressions for the estimated predicted choice probabilities for standard logit and mixed logit with only inter-individual heterogeneity can be obtained by omitting levels of integration from (23) and (24).
For each of the two prediction scenarios, we calculate Brier scores (Brier, 1950) with respect to the realised choices and the predicted choice probabilities. The Brier score (BS) of a test set is given by

$$BS = \frac{1}{NTJ} \sum_{n=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(\mathbf{1}\{y_{nt} = j\} - \widehat{P}_{ntj} \right)^2, \quad (25)$$

where $\mathbf{1}\{y_{nt} = j\}$ is an indicator, which equals one if the condition inside the braces is true and zero otherwise. $\widehat{P}_{ntj}$ is a shorthand notation for the predicted probability that $y_{nt} = j$ is observed. A lower Brier score indicates superior predictive accuracy. The Brier score is a strictly proper scoring rule, since it is exclusively minimised by the true predictive choice probabilities (Gneiting and Raftery, 2007). An important feature of the Brier score is that it takes into account the predicted choice probabilities of whole choice sets. Danaf et al. (2019) and Xie et al. (2020) use the average of the predicted probabilities of only the chosen alternatives (henceforth, P^{chosen}) to evaluate predictive accuracy, with the interpretation that a higher value of P^{chosen} indicates superior predictive performance. In what follows, we report both Brier scores and P^{chosen}.

3.3 Implementation details

We implement the MSL and MCMC estimators in Python.[1] For MSL, the numerical optimisations are performed using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Nocedal and Wright, 2006) contained in Python's SciPy library (Jones et al., 2001). Analytical gradients are provided (see Appendix A.1). The Hessian matrix of the simulated log-likelihood function is calculated as a finite difference approximation of the Jacobian of the analytical gradient. We use 250 inter-individual simulation draws per decision-maker

[1] The code is available at `https://github.com/RicoKrueger/inter_intra`.

and 250 intra-individual simulation draws per observation. To establish that the estimation results are stable for these numbers of draws, we also evaluate the performance of the MSL estimator using 500×500 draws. The simulation draws are generated using the Modified Latin Hypercube sampling (MLHS) approach (Hess et al., 2006). We also take advantage of Python's parallel processing capacities to improve the computational efficiency of the MSL estimator. We process the likelihood computations in ten parallel batches, each of which corresponds to 25 (50) inter-individual simulation draws.

The MCMC sampler for mixed logit with unobserved inter- and intra-individual heterogeneity is executed with two parallel Markov chains and 400,000 iterations for each chain, whereby the initial 200,000 iterations of each chain are discarded for burn-in. After burn-in, every tenth draw is retained to moderate storage requirements and to facilitate post-simulation computations. For standard logit and mixed logit with only inter-individual heterogeneity, the MCMC samplers are executed with two parallel Markov chains and 100,000 iterations for each chain, whereby the initial 50,000 iterations of each chain are discarded for burn-in. After burn-in, every fifth draw is kept.

3.4 Results

Table 1 compares the predictive accuracy of the models. For each value of α and number of choice situations per individual T, we report the means and the standard errors of the Brier scores as well as the average predicted probabilities of the chosen alternative (P^{chosen}) for the unconditional and the conditional prediction scenarios across 20 resamples. In our subsequent discussion, we focus on the Brier score, as it is strictly proper. Nonetheless, P^{chosen} leads to the same general conclusions.

Across the different experimental scenarios, we do not observe significant differences in unconditional predictive accuracy between the considered methods. As expected, standard logit without individual-specific parameters yields the same level of predictive accuracy in the unconditional and the conditional prediction scenarios. However, due to the presence of individual-specific parameters, mixed logit provides better conditional predictive accuracy than standard logit. For instance, in scenario 1 with $\alpha = 0.3$ for $T = 20$, MNL produces an average Brier score of 0.200. Mixed logit with only inter-individual heterogeneity produces an average Brier score of 0.152, whereas mixed logit with unobserved inter- and intra-individual heterogeneity estimated via MCMC and MSL with 250×250 draws give average Brier scores of 0.149 and 0.150, respectively. We further observe that the conditional predictive accuracy of mixed logit improves relative to standard logit, as more choice situations are included in the estimation. For example, in scenario 1 with $\alpha = 0.3$, the Brier score of mixed logit with only inter-individual heterogeneity is 0.165 for $T = 10$, while it is 0.152 for $T = 20$.

Interestingly, mixed logit with unobserved inter- and intra-individual heterogene-

ity does not provide significantly better conditional predictive accuracy than standard mixed logit in any of the considered experimental scenarios. The difference in Brier scores of the two methods is at most 0.003. Also, the proportion of variance α that is due to intra-individual taste variation does not appear to affect the conditional predictive accuracy of considered mixed logit models. For example, for $T = 20$, the average Brier score for the conditional prediction scenario of mixed logit with unobserved inter- and intra-individual heterogeneity estimated via MCMC is 0.149 and 0.150 in both scenario 1 ($\alpha = 0.3$) and scenario 2 ($\alpha = 0.7$). Even in scenario 2, in which intra-individual taste variation accounts for 70% of the total variation in tastes, mixed logit with unobserved inter- and intra-individual heterogeneity does not outperform simple mixed logit with only inter-individual heterogeneity.

Another insight from Table 1 is that the MCMC estimator and the two configurations of the MSL estimator for mixed logit with unobserved inter- and intra-individual heterogeneity perform equally well in the considered prediction scenarios.

Table 2 contrasts the estimation accuracy of MCMC estimator and the two configurations of the MSL estimator for mixed logit with unobserved inter- and intra-individual heterogeneity. We find that the three methods perform equally well at recovering parameters. We further observe negligible differences between MSL with 250×250 draws and MSL with 500×500 draws.

Finally, Table 3 gives the estimation times of the different methods across the considered experimental scenarios. Mixed logit with only inter-individual heterogeneity is substantially faster than mixed logit with unobserved inter- and intra-individual heterogeneity. In all of the considered simulation scenarios, MSL with analytical gradients and 250×250 simulation draws is faster than MCMC. For example, in scenario 1 for $T = 10$, the average estimation time of simple mixed logit is 285 seconds, while the average computation times of mixed logit with unobserved inter- and intra-individual heterogeneity estimated via MCMC and MSL with 250×250 simulation draws are approximately tenfold with 3,423 seconds and 2,951 seconds, respectively. In the considered scenarios, MSL with 250×250 draws is approximately four times faster than MSL with 500×500 draws. We also observe that the standard errors of the estimation times across the 20 resamples are proportionally higher for MSL than for MCMC. Inherent differences between the estimation algorithms explain this discrepancy. Whereas MCMC simulations are run for a fixed number of iterations, the number of function evaluations that is needed to reach convergence during the maximisation of the simulated log-likelihood is not fixed and depends on initial values and the shape of the log-likelihood surface.

In sum, the simulation study shows that none of the mixed logit models provide substantially better unconditional predictive accuracy than standard logit. Nonetheless, the considered mixed logit models offer superior conditional pre-

dictive accuracy. Yet, there are no substantive differences between standard mixed logit and a more complex mixed logit accounting for both inter- and intra-individual heterogeneity. Besides, we observe that MCMC and MSL with 250×250 draws are equivalent in terms of prediction and estimation accuracy. MSL with 250×250 draws is approximately four times faster than MSL with 500×500 draws, while offering equivalent estimation and prediction accuracy.

α	T	Method	Brier_{UC}		Brier_{C}		$\text{P}_{\text{UC}}^{\text{chosen}}$		$\text{P}_{\text{C}}^{\text{chosen}}$	
			Mean	SE	Mean	SE	Mean	SE	Mean	SE
0.3	10	MNL (MCMC)	0.20151	0.00200	0.20130	0.00229	0.39369	0.00311	0.39509	0.00372
		MXL-inter (MCMC)	0.19879	0.00222	0.16456	0.00318	0.40116	0.00336	0.54103	0.00606
		MXL-inter-intra (MCMC)	0.19847	0.00235	0.16269	0.00306	0.40380	0.00361	0.53940	0.00581
		MXL-inter-intra (MSL, 250×250 draws)	0.19836	0.00231	0.16355	0.00308	0.40323	0.00357	0.53597	0.00588
		MXL-inter-intra (MSL, 500×500 draws)	0.19837	0.00231	0.16326	0.00298	0.40375	0.00354	0.53750	0.00568
0.3	20	MNL (MCMC)	0.19910	0.00156	0.19998	0.00187	0.39785	0.00246	0.39722	0.00278
		MXL-inter (MCMC)	0.19620	0.00170	0.15204	0.00392	0.40597	0.00274	0.56657	0.00708
		MXL-inter-intra (MCMC)	0.19581	0.00181	0.14890	0.00392	0.40910	0.00296	0.56958	0.00708
		MXL-inter-intra (MSL, 250×250 draws)	0.19589	0.00177	0.14977	0.00387	0.40693	0.00294	0.56522	0.00680
		MXL-inter-intra (MSL, 500×500 draws)	0.19579	0.00182	0.14945	0.00366	0.40814	0.00302	0.56594	0.00658
0.7	10	MNL (MCMC)	0.20172	0.00206	0.20127	0.00226	0.39373	0.00339	0.39500	0.00375
		MXL-inter (MCMC)	0.19898	0.00220	0.16691	0.00293	0.40135	0.00351	0.53644	0.00542
		MXL-inter-intra (MCMC)	0.19866	0.00224	0.16397	0.00285	0.40407	0.00355	0.53597	0.00512
		MXL-inter-intra (MSL, 250×250 draws)	0.19856	0.00223	0.16424	0.00291	0.40345	0.00359	0.53215	0.00528
		MXL-inter-intra (MSL, 500×500 draws)	0.19881	0.00224	0.16496	0.00296	0.40358	0.00351	0.53375	0.00533
0.7	20	MNL (MCMC)	0.20194	0.00197	0.20583	0.00208	0.39413	0.00329	0.38792	0.00347
		MXL-inter (MCMC)	0.19960	0.00201	0.15188	0.00391	0.40106	0.00321	0.56542	0.00662
		MXL-inter-intra (MCMC)	0.19931	0.00209	0.15019	0.00380	0.40384	0.00327	0.56568	0.00636
		MXL-inter-intra (MSL, 250×250 draws)	0.19929	0.00199	0.15264	0.00381	0.40144	0.00303	0.55826	0.00645
		MXL-inter-intra (MSL, 500×500 draws)	0.19935	0.00202	0.15140	0.00365	0.40272	0.00319	0.56116	0.00635

Note: The reported values are averages and standard errors across 20 replications. α = proportion of taste variation due to intra-individual taste heterogeneity. T = observations per individual. Brier = Brier score. P^{chosen} = average predicted probability of chosen alternative. UC = unconditional prediction, C = conditional prediction.

Table 1: Predictive accuracy on simulated data

α	T	Method	RMSE(ζ) Mean	RMSE(ζ) SE	RMSE($\mathbf{\Sigma}_B$) Mean	RMSE($\mathbf{\Sigma}_B$) SE	RMSE($\mathbf{\Sigma}_W$) Mean	RMSE($\mathbf{\Sigma}_W$) SE
0.3	10	MXL-inter-intra (MCMC)	0.03744	0.00491	0.07211	0.01032	0.07137	0.00694
		MXL-inter-intra (MSL, 250×250 draws)	0.03693	0.00389	0.06640	0.00608	0.06514	0.00480
		MXL-inter-intra (MSL, 500×500 draws)	0.03379	0.00332	0.06275	0.00610	0.06374	0.00424
0.3	20	MXL-inter-intra (MCMC)	0.02319	0.00268	0.03646	0.00293	0.03903	0.00255
		MXL-inter-intra (MSL, 250×250 draws)	0.03149	0.00310	0.04556	0.00345	0.04111	0.00338
		MXL-inter-intra (MSL, 500×500 draws)	0.02612	0.00312	0.03941	0.00346	0.04004	0.00267
0.7	10	MXL-inter-intra (MCMC)	0.04355	0.00530	0.08585	0.01253	0.07898	0.00763
		MXL-inter-intra (MSL, 250×250 draws)	0.03814	0.00412	0.07142	0.00808	0.06915	0.00474
		MXL-inter-intra (MSL, 500×500 draws)	0.03889	0.00416	0.07267	0.00851	0.07004	0.00527
0.7	20	MXL-inter-intra (MCMC)	0.02119	0.00289	0.04149	0.00511	0.03965	0.00313
		MXL-inter-intra (MSL, 250×250 draws)	0.03431	0.00319	0.05014	0.00414	0.04117	0.00331
		MXL-inter-intra (MSL, 500×500 draws)	0.02798	0.00231	0.04310	0.00361	0.03950	0.00284

Note: The reported values are averages and standard errors across 20 replications. α = proportion of taste variation due to intra-individual taste heterogeneity. T = observations per individual. RMSE = root mean square error.

Table 2: Estimation accuracy on simulated data

α	T	Method	Time [s] Mean	SE
0.3	10	MNL (MCMC)	83.3	1.3
		MXL-inter (MCMC)	284.9	2.7
		MXL-inter-intra (MCMC)	3422.9	28.2
		MXL-inter-intra (MSL, 250×250 draws)	2951.0	90.5
		MXL-inter-intra (MSL, 500×500 draws)	12938.1	410.3
0.3	20	MNL (MCMC)	159.9	2.1
		MXL-inter (MCMC)	423.7	0.7
		MXL-inter-intra (MCMC)	6135.6	75.6
		MXL-inter-intra (MSL, 250×250 draws)	5214.7	185.4
		MXL-inter-intra (MSL, 500×500 draws)	23659.9	624.9
0.7	10	MNL (MCMC)	84.6	1.7
		MXL-inter (MCMC)	287.4	2.1
		MXL-inter-intra (MCMC)	3792.7	57.2
		MXL-inter-intra (MSL, 250×250 draws)	2998.4	89.5
		MXL-inter-intra (MSL, 500×500 draws)	11746.3	303.6
0.7	20	MNL (MCMC)	186.2	3.3
		MXL-inter (MCMC)	413.7	4.4
		MXL-inter-intra (MCMC)	6706.4	61.9
		MXL-inter-intra (MSL, 250×250 draws)	5323.1	129.5
		MXL-inter-intra (MSL, 500×500 draws)	22102.1	567.3

Note: The reported values are averages and standard errors across 20 replications. α = proportion of taste variation due to intra-individual taste heterogeneity. T = observations per individual.

Table 3: Estimation time on simulated data

4 Real data application

In this section, we evaluate the performance of mixed logit with unobserved inter- and intra-individual heterogeneity using real data.

4.1 Data, utility specification and implementation details

Data for the empirical application are sourced from a stated preference survey about mobility on-demand in New York City (Bansal and Daziano, 2018, Liu et al., 2018). The data include observations from 1,507 respondents who each completed seven choice situations derived from a pivot-efficient design. Each choice situation included three labelled alternatives, namely Uber (without pooling), UberPool (with pooling) and the current mode. The alternatives are described by six attributes, namely out-of-vehicle travel time (OVTT), in-vehicle travel time (IVTT), trip cost, parking cost, the powertrain of the vehicle (gas/petrol or electric) and

the automation level of the vehicle (with or without driver). Figure 1 shows an example of a choice situation.

	Uber (without ride-sharing)	**UberPool (with ride-sharing)**	**Current m**
Walking and waiting time	6 min	9 min	12 m
In-vehicle travel time	38 min	50 min	48 m
Trip cost (excl. parking cost)	\$11	\$8	\$6
Parking cost	–	–	\$6
Powertrain	Electric	Gas	Gas
Automation	Service with driver	Automated (no driver)	–

Figure 1: Example of a choice situation in the mobility-on demand experiment (reproduced from Liu et al., 2018)

Mixed logit with unobserved inter- and intra-individual heterogeneity assumes a utility specification of the following form:

$$U_{ntj} = \left(X_{ntj}^{\text{random}}\right)^{\top} \beta_{nt} + \left(X_{ntj}^{\text{fixed}}\right)^{\top} \gamma + \varepsilon_{ntj}. \tag{26}$$

Here, X_{ntj}^{random} is a vector of attributes with individual- and observation-specific random taste parameters β_{nt}, and X_{ntj}^{fixed} is a vector of attributes with fixed taste parameters γ. ε_{ntj} is a stochastic disturbance with distribution Gumbel$(0, 1)$.
We performed an extensive specification search to determine which attributes to associate with either random or fixed parameters. During the specification search, we monitored model tractability and the inferred amounts of intra-individual taste variation. In the final model specification, we include three random parameters and four fixed parameters in the model. The random parameters pertain to OVTT, IVTT and a dummy variable indicating whether the hypothetical mobility on-demand vehicle is automated. The fixed parameters pertain to two alternative-specific constants for the hypothetical mobility on-demand alternatives, a dummy variable indicating whether the hypothetical mobility on-demand vehicle is electric and the total trip cost which subsumes the trip cost and the parking cost. The dummy variables are effects-coded with negative one indicating that the feature is absent and positive one indicating that the feature is present to reduce the scale of the associated parameters. OVTT and IVTT are divided by ten to increase the scale of the associated parameters. The utilities of standard logit and mixed logit with only inter- and intra-individual heterogeneity are specified analogously.
We consider two configurations of the training and the test data to evaluate the influence of including different numbers of choice situations per individual in the training data on the predictive accuracy. In the first configuration, the training set includes four randomly selected choice situations from 1,407 randomly selected respondents. One test set is used to evaluate the unconditional predictive ability of the considered models. It includes one choice situation from each of the remaining 100 respondents. A second test set is used to evaluate the conditional predictive

ability. It is formed by randomly selecting one of the remaining choice situations from the 1,407 respondents included in the training sample. In the second configuration, the training set includes six randomly selected choice situations from 1,407 randomly selected respondents. The test sets are created in the same way as in the first configuration. For each configuration, we create ten random splits into training and test sets. We then compare the performance of the different choice models across the splits.
The MCMC methods are estimated in the same way as described in Section 3.3. For MSL, we use 250 inter-individual simulation draws per decision-maker and 250 intra-individual simulation draws per observation. We also tested the MSL method with 500×500 draws but found no differences in prediction accuracy and estimation results.

4.2 Results

Table 4 contrasts the predictive accuracy of the considered models. For each configuration of the training data with $T = 4$ or $T = 6$ and for each model, we report the means and the standard errors of the Brier scores and the average predicted choice probabilities of the chosen alternative for the unconditional and the conditional prediction scenarios across ten random splits of the training data. Overall, the results are consistent with the results of the simulation study. We do not observe any noteworthy differences in unconditional predictive accuracy across methods and configurations of the training data. Both mixed logit models offer better conditional prediction accuracy than the standard multinomial logit model. The more complex mixed logit model with unobserved inter- and intra-individual heterogeneity does not provide benefits over standard mixed logit in terms of conditional prediction accuracy. For both types of mixed logit, the conditional prediction accuracy increases as more choice situations are included in the training data. For example, for $T = 4$, standard mixed logit produces an average Brier score of 0.154 in the conditional prediction scenarios. The same model yields an average Brier score of 0.147 in the conditional prediction scenarios with $T = 6$. In both configurations of the training data, the MCMC and MSL estimators for mixed logit with unobserved inter- and intra-individual heterogeneity are equally accurate in the two prediction scenarios.

T	Method	Brier_{UC} Mean	SE	Brier_{C} Mean	SE	p^{chosen}_{UC} Mean	SE	p^{chosen}_{C} Mean	SE
4	MNL (MCMC)	0.17925	0.00478	0.18290	0.00070	0.45512	0.00751	0.44997	0.00132
	MXL-inter (MCMC)	0.18005	0.00511	0.15376	0.00065	0.45387	0.00807	0.56771	0.00181
	MXL-inter-intra (MCMC)	0.18079	0.00524	0.15131	0.00060	0.45288	0.00820	0.57228	0.00163
	MXL-inter-intra (MSL)	0.18066	0.00523	0.15163	0.00061	0.45349	0.00818	0.57064	0.00172
6	MNL (MCMC)	0.18989	0.00405	0.18165	0.00066	0.44018	0.00616	0.45197	0.00128
	MXL-inter (MCMC)	0.18860	0.00396	0.14694	0.00161	0.44263	0.00629	0.58253	0.00233
	MXL-inter-intra (MCMC)	0.18947	0.00382	0.14483	0.00155	0.44131	0.00604	0.58804	0.00232
	MXL-inter-intra (MSL)	0.18940	0.00381	0.14513	0.00161	0.44144	0.00601	0.58665	0.00258

Note: The reported values are averages and standard errors across ten random splits. T = observations per individual. Brier = Brier score. p^{chosen} = average predicted probability of chosen alternative. UC = unconditional prediction, C = conditional prediction.

Table 4: Predictive accuracy on real data

Table 5 enumerates the detailed estimation results for one of the random splits of the stated choice data with six choice situations per individual. For MCMC, we report the posterior means, the posterior standard deviations and the bounds of the 95% credible interval. For MSL, we report the point estimates, the asymptotic standard errors and the bounds of the 95% confidence intervals. Recall that for MSL, the maximisation of the simulated log-likelihood is performed with respect to the Cholesky factors of the covariance matrices of the heterogeneity distributions. Thus, the reported estimates of the covariance elements reported are derived from the point estimates of the Cholesky factors. Standard errors of the covariance elements are obtained using a parametric bootstrap with 10,000 draws.

Our first observation is that in the majority of cases, the fixed taste parameters and means of the random taste parameters have the same signs in the four models. All four models suggest that the fixed taste parameter with respect to vehicle electrification is not statistically different. In the case of the mean of the random taste parameter pertaining to OVTT, we observe that the parameter is negative and statistically different from zero in standard mixed logit, while the parameter is not statistically different from zero in both classical and Bayesian mixed logit with unobserved inter- and intra-individual heterogeneity. Furthermore, we find that the MSL and MCMC estimates, including the credible (confidence) intervals, of mixed logit with unobserved inter- and intra-individual heterogeneity exhibit a close correspondence.

Due to its ability to decompose taste variation into inter- and intra-individual components, mixed logit with unobserved inter- and intra-individual heterogeneity offers interesting behavioural insights into the sources of taste variation. We find evidence of substantial intra-individual taste variation. For example, MCMC suggests that $2.169/(3.782+2.169) = 36.4\%$ of the variation in tastes with respect to OVTT are due to intra-individual heterogeneity. Similarly, MSL indicates that $0.644/(1.260 + 0.644) = 33.8\%$ of the variation in tastes with respect to vehicle automation can be ascribed to intra-individual heterogeneity.

Both MCMC and MSL suggest that all off-diagonal elements of the inter-individual

covariance of mixed logit with unobserved inter- and intra-individual heterogeneity are statistically different from zero. Also, both MCMC and MSL suggest that with the exception of the covariance between the random parameters pertaining to vehicle automation and IVTT, all off-diagonal elements of the intra-individual covariance of mixed logit with unobserved inter- and intra-individual heterogeneity are statistically different from zero.

Finally, Table 6 gives the estimation times of the models across the ten random splits for the two configurations of the training data. We observe that mixed logit with only inter-individual heterogeneity is substantially faster than mixed logit with unobserved inter- and intra-individual heterogeneity. For example, for $T = 4$, the average estimation time of standard mixed logit estimated via MCMC is approximately seven times lower than the average estimation time of mixed logit with unobserved inter- and intra-individual heterogeneity estimated via MCMC and approximately eleven times lower than the average estimation time of mixed logit with unobserved inter- and intra-individual heterogeneity estimated via MSL.

Parameter	MNL (MCMC)				MXL-inter (MCMC)				MXL-inter-intra (MCMC)				MXL-inter-intra (MSL)			
	Est.	SE	[2.5%	97.5%]	Est.	SE	[2.5%	97.5%]	Est.	SE	[2.5%	97.5%]	Est.	SE	[2.5%	97.5%]
Fixed parameters																
ASC Uber	-0.576	0.051	-0.647	-0.459	-0.569	0.076	-0.725	-0.405	-0.600	0.086	-0.767	-0.438	-0.606	0.087	-0.778	-0.435
ASC UberPool	-1.045	0.024	-1.124	-1.019	-1.021	0.056	-1.128	-0.909	-1.095	0.065	-1.231	-0.966	-1.100	0.066	-1.229	-0.971
Total trip cost	-0.072	0.003	-0.080	-0.067	-0.103	0.005	-0.114	-0.094	-0.121	0.007	-0.136	-0.107	-0.121	0.007	-0.135	-0.107
Electric	-0.031	0.026	-0.086	0.020	0.016	0.026	-0.037	0.064	0.020	0.032	-0.042	0.084	0.019	0.032	-0.044	0.082
OVTT	-0.329	0.021	-0.353	-0.294												
IVTT	-0.240	0.019	-0.261	-0.189												
Automated	-0.131	0.027	-0.173	-0.078												
Random parameters: Means																
OVTT					-0.168	0.055	-0.274	-0.057	0.020	0.095	-0.160	0.211	-0.022	0.095	-0.209	0.164
IVTT					-0.474	0.048	-0.569	-0.382	-0.585	0.067	-0.722	-0.464	-0.562	0.063	-0.686	-0.439
Automated					-0.392	0.040	-0.473	-0.316	-0.665	0.116	-0.929	-0.464	-0.630	0.096	-0.818	-0.442
Random parameters: Inter-individual covariance																
OVTT vs. OVTT					1.251	0.167	0.953	1.610	3.782	0.759	2.502	5.396	3.386	0.644	2.258	4.794
IVTT vs. OVTT					-0.236	0.077	-0.386	-0.086	-0.453	0.163	-0.808	-0.166	-0.322	0.153	-0.627	-0.029
IVTT vs. IVTT					0.576	0.097	0.399	0.776	0.930	0.202	0.632	1.393	0.928	0.164	0.642	1.285
Automated vs. OVTT					-0.613	0.070	-0.755	-0.482	-1.633	0.364	-2.463	-1.056	-1.527	0.300	-2.160	-0.983
Automated vs. IVTT					0.158	0.051	0.062	0.261	0.207	0.101	0.030	0.433	0.214	0.092	0.044	0.405
Automated vs. automated					0.725	0.073	0.591	0.878	1.372	0.315	0.872	2.124	1.260	0.241	0.850	1.805
Random parameters: Intra-individual covariance																
OVTT vs. OVTT									2.169	0.561	1.238	3.403	2.072	0.520	1.188	3.213
IVTT vs. OVTT									0.246	0.146	0.014	0.569	0.390	0.167	0.082	0.742
IVTT vs. IVTT									0.119	0.079	0.021	0.329	0.127	0.077	0.029	0.325
Automated vs. OVTT									0.383	0.159	0.102	0.736	0.404	0.156	0.132	0.743
Automated vs. IVTT									-0.017	0.091	-0.243	0.143	-0.098	0.090	-0.280	0.076
Automated vs. automated									0.709	0.346	0.167	1.525	0.644	0.372	0.314	1.748

Note: The dummy variables electric and automated are effects-coded with negative one indicating that the feature is absent and positive one indicating that the feature is present to reduce the scale of the associated parameters. OVTT and IVTT are divided by ten to increase the scale of the associated parameters.

Table 5: Estimation results for one of the random splits of the real data with six choice situations per individual

T	Method	Time [s] Mean	SE
4	MNL (MCMC)	60.9	0.2
	MXL-inter (MCMC)	344.1	6.4
	MXL-inter-intra (MCMC)	2434.6	75.8
	MXL-inter-intra (MSL)	3754.5	332.2
6	MNL (MCMC)	85.1	1.8
	MXL-inter (MCMC)	399.1	13.2
	MXL-inter-intra (MCMC)	3147.0	55.2
	MXL-inter-intra (MSL)	6027.0	401.8

Note: The reported values are averages and standard errors across ten random splits. T = observations per individual.

Table 6: Estimation time on real data

5 Extended discussion

Our analysis suggests that mixed logit models with unobserved inter- and intra-individual heterogeneity do not provide significant improvements over simpler mixed logit models which only account for unobserved inter-individual heterogeneity in terms of conditional prediction accuracy. The inability of the former to outperform the latter can be ascribed to the former's predominant emphasis on nonstructural random heterogeneity. Thus, there is a need to explore alternative modelling approaches which have the potential to provide accurate individualised predictions of choice behaviour by accounting for richer dependencies between products and consumers' preferences as well as temporal correlations between choices in a flexible framework. In what follows, we discuss four strands of the literature and evaluate their relevance in creating choice-based recommender systems within the random utility maximisation (RUM) framework.

5.1 Collaborative filtering

Various collaborative filtering approaches such as matrix factorisation have emerged as powerful tools to generate personalised recommendations in recommender systems (Gopalan et al., 2013, Koren et al., 2009, Mnih and Salakhutdinov, 2008). The fundamental idea of collaborative filtering is to predict a consumer's preferences by exploiting interdependencies between products. Matrix factorisation provides a mapping of consumers and products into a joint latent factor space. Matrix factorisation consists of learning a sparse matrix of dimension # of consumers $\times$ # of products. Each cell of this matrix represents one consumer's preference for a specific product. The consumer's preference is represented as the inner product

of a latent vector of product characteristics and a latent vector of consumer preferences for each of the latent product characteristics (see Gopalan et al., 2013, for details of the formulation).
Learning such a sparse matrix is computationally challenging, but advancements in variational Bayes have made the estimation of these models tractable for large data sets. Recent studies on matrix factorisation methods also account for dynamic consumer preferences and social network effects (Hosseini et al., 2018). A combination of scalability, ability to account for dynamics and social aspects, and superior predictive accuracy have made matrix factorisation methods popular in industrial applications. However, they have received limited attention of applied econometrics and marketing communities due to their i) predominant focus on prediction rather than inference, ii) apparent disconnection to economic theory, iii) inability to model time-varying choice sets and product-specific attributes. Economists and machine learning researchers recently joined forces to address the second and third limitations of this powerful tool. Athey et al. (2018) illustrate how matrix factorisation methods can be integrated into standard RUM frameworks to predict an individual's choice of restaurants using data from local search-and-discovery application. The main idea of the approach is to augment the original utility equation with the consumer- and product-level covariates by including a vector of latent characteristics for each restaurant as well as latent preferences of consumers for these characteristics. Thus, the framework incorporates the key component (i.e., sparse latent construct) of standard matrix factorisation models in the RUM framework and adopts variational Bayes for scalable estimation and prediction. In another study, Donnelly et al. (2019) use a similar framework to model consumer preferences across multiple categories of products in a supermarket. These theory-driven advancements would hopefully convince applied choice modellers about the benefits of matrix factorisation methods for personalised predictions.

5.2 Collaborative learning

Zhu et al. (2020) propose a choice model with time-varying parameters in a collaborative learning framework. Similar to latent class models, this model assumes that there are several unique underlying preference patterns (i.e., classes), but rather than assigning each consumer to one class and assuming preferences of all class members to be the same, a vector of weights (membership vector) is specified to represent the degree of resemblance of the consumer's preferences to each preference pattern. Temporal variation in these unique preference patterns is captured by time-varying model parameters. The framework is already a viable alternative to mixed logit with inter-and intra-heterogeneity. Yet, it be can further improved by taking inspiration from Athey et al. (2018) and by incorporating the latent structure of matrix factorisation into the utility equation.

5.3 Amortised variational inference

Recent application of amortised variational inference (AVI) in the estimation of the mixed logit model also offers possibilities to improve the choice prediction accuracy (Rodrigues, 2020). Instead of introducing consumer-level local variational parameters for random parameters, AVI maps observed choices and covariates with corresponding variational parameters using a deep neural network to avoid the growth of variational parameters with the sample size. AVI thus includes a generic inference network that takes a consumer's data as input and provides the approximate posterior distribution of her random taste parameters as output. In other words, AVI provides a trained inference network as a byproduct of the estimation, which can be used to obtain the posterior distribution of random taste parameters of a new consumer or the existing consumer in a new choice situation (Rodrigues, 2020).
AVI has the potential to become a workhorse method in online learning applications due to its fast estimation with stochastic backpropagation and GPU-accelerated computations. AVI performs well in the initial experiments presented in Rodrigues (2020), but its performance needs to be benchmarked against other competing methods.

5.4 Neural network and tree-based models

To leverage benefits of machine learning advancements in discrete choice models without compromising at interpretability and economic theory, recent RUM based choice models have adopted variants of neural networks (Sifringer et al., 2020, Wang et al., 2020) and regression trees (Kindo et al., 2016) to specify semi- and non-parametric utility functions. These advanced models claim to improve the prediction accuracy of discrete choice models in validation samples, but they have limited focus on improving within individual predictions, i.e. predicting choice of a consumer from training dataset in a new choice situation. Bringing this additional feature in these data-theory-driven models can make them viable for online recommender systems.

6 Conclusion

In this research note, we evaluate the ability of mixed logit models with unobserved inter- and intra-individual heterogeneity to generate individual-level predictions. Using simulated and real data, we demonstrate that mixed logit with unobserved inter- and intra-individual heterogeneity does not provide significant improvements over standard mixed logit models which only account for inter-individual taste variation. This observation persists even in scenarios which are

characterised by high levels of intra-individual taste variation and when the number of choice situations per individual is large.
Besides, the estimation of mixed logit with unobserved inter- and intra-individual heterogeneity demands at least seven times as much computation time as the estimation of standard mixed logit. For mixed logit with unobserved inter- and intra-individual heterogeneity, we also find that the maximum simulated likelihood (MSL) estimator with analytical gradients is faster or not substantially slower then the Bayesian Markov chain Monte Carlo (MCMC) method, which stands in contrast to previous studies which used MSL with numerical gradients (see Becker et al., 2018).
We ascribe the inability of mixed logit with unobserved inter- and intra-individual heterogeneity to outperform standard mixed logit to the former's predominant emphasis on nonstructural random heterogeneity. In light of recent advances at the intersection of machine learning and econometrics, we discussed several promising alternative modelling approaches, which may offer superior prediction performance by flexibly capturing dependencies between decision-makers, alternatives and attributes.
Regardless of our findings, we argue that mixed logit with unobserved inter- and intra-individual heterogeneity has a place in the literature as a variance decomposition technique and for understanding behaviour, even though the model does not offer substantially better predictive accuracy than standard mixed logit models which only account for inter-individual heterogeneity. Whilst extant applications of the model focus on decomposing taste variation into inter- and intra-individual terms, future applications of the model may benefit from reconceptualising the model as an instance of a hierarchical model (Gelman and Hill, 2006) and may examine less granular nesting structures of observational units. In particular, revealed preference data inherently provide many meaningful ways to organise observational units into nests. For example, in a longitudinal study of the career choices of high school graduates, the observational units are naturally nested within schools, districts, graduation year etc.
Our analysis also includes comparisons of the unconditional out-of-sample predictive abilities of standard logit, standard mixed logit and mixed logit with unobserved inter- and intra-individual heterogeneity. In both the simulation study and the real data application, we find that the two types of mixed logit do not offer substantial improvements in unconditional predictive accuracy over standard logit. These observations are consistent with the literature (Cherchi and Cirillo, 2010, Wang et al., 2021, Zhao et al., 2020) and suggest that while mixed logit is useful for explaining behaviour, as it can accommodate unobserved taste heterogeneity, unrestricted substitution patterns and correlation in unobservables over time, mixed logit may not necessarily provide more accurate unconditional out-of-sample predictions than standard logit.
Finally, we note that the out-of-sample prediction scenarios considered in this

research note correspond to cases of internal validation (see Parady et al., 2021). It would be instructive to assess to what extent our findings generalise to cases of external validation, i.e. out-of-sample prediction using data from a different time period or region or data gathered via a different method (see Parady et al., 2021).

References

Akinc, D. and Vandebroek, M. (2018). Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix. *Journal of choice modelling*, 29:133–151.

Ansari, A., Essegaier, S., and Kohli, R. (2000). Internet recommendation systems.

Athey, S., Blei, D., Donnelly, R., Ruiz, F., and Schmidt, T. (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. In *AEA Papers and Proceedings*, volume 108, pages 64–67.

Bansal, P. and Daziano, R. A. (2018). Influence of choice experiment designs on eliciting preferences for autonomous vehicles. *Transportation Research Procedia*, 32:474–481.

Bansal, P., Krueger, R., Bierlaire, M., Daziano, R. A., and Rashidi, T. H. (2020). Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. *Transportation Research Part B: Methodological*, 131:124 – 142.

Becker, F., Danaf, M., Song, X., Atasoy, B., and Ben-Akiva, M. (2018). Bayesian estimator for logit mixtures with inter-and intra-consumer heterogeneity. *Transportation Research Part B: Methodological*, 117:1–17.

Ben-Akiva, M., McFadden, D., Train, K., et al. (2019). Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis. *Foundations and Trends® in Econometrics*, 10(1-2):1–144.

Bettman, J. R., Luce, M. F., and Payne, J. W. (1998). Constructive consumer choice processes. *Journal of consumer research*, 25(3):187–217.

Bhat, C. R. and 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. *Transportation Research Part B: Methodological*, 36(7):593–616.

Bhat, C. R. and Sardesai, R. (2006). The impact of stop-making and travel time reliability on commute mode choice. *Transportation Research Part B: Methodological*, 40(9):709–730.

Bhat, C. R. and Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (macml) estimator for mixed multinomial probit models. *Transportation Research Part B: Methodological*, 45(7):940–953.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Cherchi, E. and Cirillo, C. (2010). Validation and forecasts in models estimated from multiday travel survey. *Transportation research record*, 2175(1):57–64.

Danaf, M., Becker, F., Song, X., Atasoy, B., and Ben-Akiva, M. (2019). Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems*, 119:35–45.

Donnelly, R., Ruiz, F. R., Blei, D., and Athey, S. (2019). Counterfactual inference for consumer choice across many product categories. *arXiv preprint arXiv:1906.02635*.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2013). Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*.

Hess, S. and Giergiczny, M. (2015). Intra-respondent heterogeneity in a stated choice survey on wetland conservation in belarus: first steps towards creating a link with uncertainty in contingent valuation. *Environmental and Resource Economics*, 60(3):327–347.

Hess, S. and 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. and Train, K. E. (2011). Recovery of inter-and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological*, 45(7):973–990.

Hess, S., Train, K. E., and Polak, J. W. (2006). On the use of a modified latin hypercube sampling (mlhs) method in the estimation of a mixed logit model for vehicle choice. *Transportation Research Part B: Methodological*, 40(2):147–163.

Hosseini, S. A., Khodadadi, A., Alizadeh, K., Arabzadeh, A., Farajtabar, M., Zha, H., and Rabiee, H. R. (2018). Recurrent poisson factorization for temporal recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(1):121–134.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.*, 8(2):439–452.

Jiang, H., Qi, X., and Sun, H. (2014). Choice-based recommender systems: a unified approach to achieving relevancy and diversity. *Operations Research*, 62(5):973–993.

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

Kim, S. (2015). How Foursquare and Other Apps Guess What You Want to Eat.

Kindo, B. P., Wang, H., and Peña, E. A. (2016). Multinomial probit bayesian additive regression trees. *Stat*, 5(1):119–131.

Kivetz, R., Netzer, O., and Schrift, R. (2008). The synthesis of preference: Bridging behavioral decision research and marketing science. *Journal of Consumer Psychology*, 18(3):179–186.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.

Liu, Y., Bansal, P., Daziano, R., and Samaranayake, S. (2018). A framework to integrate mode choice in the design of mobility-on-demand systems. *Transportation Research Part C: Emerging Technologies*.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems*, 74:12–32.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470.

Mnih, A. and Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Parady, G., Ory, D., and Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 38:100257.

Revelt, D. and Train, K. (1998). Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *The Review of Economics and Statistics*, 80(4):647–657.

Revelt, D. and Train, K. (2000). Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier.

Rodrigues, F. (2020). Scaling bayesian inference of mixed multinomial logit models to very large datasets. *arXiv preprint arXiv:2004.05426*.

Sifringer, B., Lurkin, V., and Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140:236–261.

Song, X., Danaf, M., Atasoy, B., and Ben-Akiva, M. (2018). Personalized menu optimization with preference updater: a boston case study. *Transportation Research Record*, 2672(8):599–607.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The american economic review*, 67(2):76–90.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition.

Wang, S., Mo, B., Hess, S., and Zhao, J. (2021). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: an empirical benchmark. *arXiv preprint arXiv:2102.01130*.

Wang, S., Mo, B., and Zhao, J. (2020). Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies*, 112:234–251.

Xie, Y., Zhang, Y., Akkinepally, A. P., and Ben-Akiva, M. (2020). Personalized choice model for managed lane travel behavior. *Transportation research record*, 2674(7):442–455.

Yáñez, M. F., Cherchi, E., Heydecker, B. G., and de Dios Ortúzar, 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.

Zhao, X., Yan, X., Yu, A., and Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society*, 20:22–35.

Zhu, X., Feng, J., Huang, S., and Chen, C. (2020). An online updating method for time-varying preference learning. *Transportation Research Part C: Emerging Technologies*, 121:102849.

A Estimation details

A.1 Gradient of simulated log-likelihood

In what follows, we derive expressions for the gradients of the simulated log-likelihood of mixed logit model with unobserved inter- and intra-individual heterogeneity. First, we let ϑ_i denote one of the model parameters collected in θ. We have

$$\frac{\partial}{\partial \vartheta_i} SLL(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \frac{\frac{1}{D}\sum_{d=1}^{D} \frac{\partial}{\partial \vartheta_i} \prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \right)}{\frac{1}{D}\sum_{d=1}^{D} \prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \right)}. \tag{27}$$

To find the derivative in the numerator, we define

$$\psi_{nt,d}(\theta) = \frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \tag{28}$$

with

$$\begin{aligned}
\psi'_{nt,d}(\theta) =& \frac{\partial \psi_{nt,dr}(\boldsymbol{\Theta})}{\partial \vartheta_i} \\
=& \frac{1}{R} \sum_{r=1}^{R} \left(P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \frac{\partial V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt,dr})}{\partial \vartheta_i} \right. \\
& \left. - \sum_{j' \in \mathcal{C}: j' \neq y_{nt}} \left(P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) P(j'|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \frac{\partial V(\mathbf{X}_{ntj'}, \boldsymbol{\beta}_{nt,dr})}{\partial \vartheta_i} \right) \right).
\end{aligned} \tag{29}$$

Note that if the deterministic aspect of the utility is specified as linear-in-parameters, i.e.

$$V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt,dr}) = \mathbf{X}_{ntj}^{\top} (\zeta + \mathbf{L}_B \boldsymbol{\xi}_{n,d} + \mathbf{L}_W \boldsymbol{\xi}_{nt,r}), \tag{30}$$

we have $\frac{\partial V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt,dr})}{\partial \zeta} = \mathbf{X}_{ntj}$, $\frac{\partial V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt,dr})}{\partial \mathbf{L}_B} = \mathbf{X}_{ntj} \boldsymbol{\xi}_{n,d}^{\top}$, and $\frac{\partial V(\mathbf{X}_{ntj}, \boldsymbol{\beta}_{nt,dr})}{\partial \mathbf{L}_W} = \mathbf{X}_{ntj} \boldsymbol{\xi}_{nt,r}^{\top}$. From the product rule of differentiation, it follows that

$$\frac{\partial}{\partial \vartheta_i} \prod_{t=1}^{T} \left(\frac{1}{R} \sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr}) \right) = \left(\prod_{t=1}^{T} \psi_{nt,d}(\theta) \right) \left(\sum_{t=1}^{T} \frac{\psi'_{nt,d}(\theta)}{\psi_{nt,d}(\theta)} \right). \tag{31}$$

A.2 Gibbs sampler

In what follows, we present one iteration of the Gibbs sampler for mixed logit with unobserved inter- and intra-individual heterogeneity:

1. Update $\delta_{B,k}$ for all $k \in \{1, \ldots, K\}$ by sampling $\delta_{B,k} \sim \text{Gamma}\left(\frac{\nu_B + K}{2}, \frac{1}{a_{B,k}^2} + \nu_B \left(\boldsymbol{\Sigma}_B^{-1} \right)_{kk} \right)$.

2. Update $\boldsymbol{\Sigma}_B$ by sampling $\boldsymbol{\Sigma}_B \sim \text{IW}\Big(\nu_B + N + K - 1, 2\nu_B \text{diag}(\boldsymbol{\delta}_B) + \sum_{n=1}^{N}(\boldsymbol{\mu}_n - \boldsymbol{\zeta})(\boldsymbol{\mu}_n - \boldsymbol{\zeta})^\top\Big)$.

3. Update $\delta_{W,k}$ for all $k \in \{1, \ldots, K\}$ by sampling $\delta_{W,k} \sim \text{Gamma}\Big(\frac{\nu_W + K}{2}, \frac{1}{A_{W,k}^2} + \nu_W \left(\boldsymbol{\Sigma}_W^{-1}\right)_{kk}\Big)$.

4. Update $\boldsymbol{\Sigma}_W$ by sampling $\boldsymbol{\Sigma}_W \sim \text{IW}\Big(\nu_W + \sum_{n=1}^{N} T + K - 1, 2\nu_W \text{diag}(\boldsymbol{\delta}_W) + \sum_{n=1}^{N}\sum_{t=1}^{T}(\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)(\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)^\top\Big)$.

5. Update $\boldsymbol{\zeta}$ by sampling $\boldsymbol{\zeta} \sim N(\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$, where $\boldsymbol{\Sigma}_\zeta = \left(\boldsymbol{\Lambda}_0^{-1} + N\boldsymbol{\Sigma}_B^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_\zeta = \boldsymbol{\Sigma}_\zeta\Big(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\lambda}_0 + \boldsymbol{\Sigma}_B^{-1}\sum_{n=1}^{N}\boldsymbol{\mu}_n\Big)$.

6. Update $\boldsymbol{\mu}_n$ for all $n \in \{1, \ldots, N\}$ by sampling $\boldsymbol{\mu}_n \sim N(\boldsymbol{\mu}_{\mu_n}, \boldsymbol{\Sigma}_{\mu_n})$, where $\boldsymbol{\Sigma}_{\mu_n} = \left(\boldsymbol{\Sigma}_B^{-1} + T\boldsymbol{\Sigma}_W^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\mu_n} = \boldsymbol{\Sigma}_{\mu_n}\left(\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\zeta} + \boldsymbol{\Sigma}_W^{-1}\sum_{t=1}^{T}\boldsymbol{\beta}_{nt}\right)$.

7. Update $\boldsymbol{\beta}_{nt}$ for all $n \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$:
   (a) Propose $\tilde{\boldsymbol{\beta}}_{nt} = \boldsymbol{\beta}_{nt} + \sqrt{\rho}\,\text{chol}(\boldsymbol{\Sigma}_W)\boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{I}_K)$.
   (b) Compute $r = \frac{P(y_{nt}|\mathbf{X}_{nt}, \tilde{\boldsymbol{\beta}}_{nt})\phi(\tilde{\boldsymbol{\beta}}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)}{P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt})\phi(\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)}$.
   (c) Draw $u \sim \text{Uniform}(0, 1)$. If $r \leq u$, accept the proposal. If $r > u$, reject the proposal.

ρ is a step size, which needs to be tuned. We employ the same tuning mechanism as Train (2009): ρ is set to an initial value of 0.1 and after each iteration, ρ is decreased by 0.001, if the average acceptance rate across all decision-makers is less than 0.3; ρ is increased by 0.001, if the average acceptance rate across all decision-makers is more than 0.3.

B True population parameters in the simulation study

$$\boldsymbol{\zeta} = \begin{bmatrix} -0.5 & 0.5 & -0.5 & 0.5 \end{bmatrix}^\top, \boldsymbol{\Omega}_B = \mathbf{I}_4 + \alpha \cdot \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \boldsymbol{\Omega}_W = \mathbf{I}_4 + \alpha \cdot \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$