On the usefulness of mixed logit models with unobserved inter- and intra-individual heterogeneity

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

December 8, 2020

Report TRANSP-OR 201208
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 as well as across choice tasks encountered by the same individual. Recent work advocates the use of 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. In this research note, 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 with unobserved inter- and intra-individual heterogeneity does 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 ten times as much computation time as the estimation of standard mixed logit models. Informed by recent advances in machine learning and econometrics, we then discuss alternative modelling approaches, which can capture richer dependencies between decision-makers, alternatives and attributes.

Keywords: mixed logit; unobserved heterogeneity; personalised recommendations.

1 Introduction

The representation of taste heterogeneity is a principal concern of discrete choice analysis, as information on the distribution of tastes is critical for demand forecasting, welfare analysis and market segmentation (e.g. Allenby and Rossi, 1998, Ben-Akiva et al., 2019). From the analyst's perspective, taste variation is often random because differences in sensitivities cannot be related to observed or observable characteristics of the decision-maker or features of the choice context (see e.g Bhat, 1998, 2000).

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). 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 the use of 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 between (inter-individual prediction for respondents without a history of past choices) and within (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 account for inter-individual heterogeneity.[1]

With the growing availability of dynamic panel data sets, recommender systems are increasingly employed to tailor recommendations of goods and services to individual-specific preferences (Ansari et al., 2000, Lu et al., 2015). Recommender systems increase

[1] Danaf et al. (2019) and Xie et al. (2020) compare personalised and non-personalised predicted choice probabilities of mixed logit with inter- and intra-individual heterogeneity. They conclude that personalisation improves within-individuals predictive accuracy. However, non-personalised choice probabilities of mixed logit with inter- and intra-individual heterogeneity are not the same as personalised choice probabilities of standard mixed logit.

user satisfaction and lower search costs by helping users to navigate intricate choice sets in complex goods and service services systems such as Internet marketplaces and smart mobility (Ansari et al., 2000, Song et al., 2018). 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). By accounting for alternative-specific attributes, discrete choice models also better capture product diversity (Jiang et al., 2014).

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 within-individuals 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 to generate personalised predictions with random utility models.

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 a simulation evaluation and a 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

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 $t \in \{1, \ldots T\}$, a decision-maker $n \in \{1, \ldots N\}$ derives utility

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

from alternative j in the set $\mathcal{C} = \{1, \ldots, J\}$. 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. The assumption $\varepsilon_{ntj} \overset{\text{iid}}{\sim} \text{Gumbel}(0, 1)$ leads to the logit model such that 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 $y_{nt} \in \mathcal{C}$ is an indicator of the observed choice.

Note that in equation (1), the taste parameters $\boldsymbol{\beta}_{nt}$ are defined as being observation-specific. To allow for dependence between repeated observations for the same individual and to accommodate inter-individual taste heterogeneity, it has become standard practice to adopt Revelt's and Train's (1998) panel estimator of mixed logit. Under this specification, taste homogeneity across replications is assumed such that $\boldsymbol{\beta}_{nt} = \boldsymbol{\beta}_n$ $\forall t \in \{1, \ldots, T\}$. To also accommodate intra-individual taste heterogeneity in addition

to inter-individual taste heterogeneity, the taste vector $\boldsymbol{\beta}_{nt}$ can be defined as 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 covariance matrix. The distribution of individual-specific parameters $\boldsymbol{\mu}_{1:N}$ is then also assumed to be multivariate normal, i.e. $\boldsymbol{\mu}_n \sim N(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B)$ for $n = 1, \dots, N$, where $\boldsymbol{\zeta}$ is a mean vector and $\boldsymbol{\Sigma}_B$ is a covariance matrix.
Mixed logit with unobserved inter- and intra-individual heterogeneity can be estimated using either maximum simulated likelihood (MSL) or Bayesian Markov chain Monte Carlo (MCMC) methods. In Appendix A, we describe both estimation approaches. Whereas in MSL, the individual-specific parameters $\boldsymbol{\mu}_n$ are treated as stochastic nuisance parameters, these parameters are directly estimated in Bayesian estimation. Thus, it is particularly easy to make individual-specific predictions with Bayesian methods.

3 Simulation study

In this section, we present an extensive simulation evaluation of mixed logit with unobserved inter- and intra-individual heterogeneity. We benchmark the performance of the model in terms of estimation time, estimation accuracy and out-of-sample predictive accuracy against simpler standard logit and mixed logit models with only inter-individual heterogeneity. We also contrast the performance of the MSL and MCMC estimators of mixed logit with unobserved inter- and intra-individual heterogeneity in terms of estimation time and accuracy.

3.1 Data and experimental setup

For the simulation study, we rely on synthetic choice data, which we create 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

$$U_{ntj} = \mathbf{X}_{ntj}^\top \boldsymbol{\beta}_{nt} + \varepsilon_{ntj}. \tag{3}$$

The definition of the variables is the same as in Section 2.
We consider two experimental scenarios with different proportions of total variance that be ascribed to intra-individual taste variation for the generation of the case-specific taste parameters $\boldsymbol{\beta}_{nt}$. In both scenarios, $\boldsymbol{\beta}_{nt}$ are drawn via the following process:

$$\boldsymbol{\mu}_n | \boldsymbol{\zeta}, \boldsymbol{\Sigma}_B \sim N(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_B), n = 1, \dots, N, \tag{4}$$

$$\boldsymbol{\beta}_{nt} | \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W), n = 1, \dots, N, t = 1, \dots, T, \tag{5}$$

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 $\boldsymbol{\zeta}$, $\boldsymbol{\Omega}_B$ and $\boldsymbol{\Omega}_W$ are enumerated in Appendix B. We define $\boldsymbol{\sigma}_B^2 = 2 \cdot (1 - \alpha) \cdot |\boldsymbol{\zeta}|$ and $\boldsymbol{\sigma}_W^2 = 2 \cdot \alpha \cdot |\boldsymbol{\zeta}|$ 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. We vary α across the two scenarios: In scenario 1, we let $\alpha = 0.3$; in scenario 2, we let $\alpha = 0.7$.

In both scenarios, the alternative-specific attributes 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 deterministically-best alternative due to the stochastic utility component. In each scenario, we 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 sample and in terms of their 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 $\{\boldsymbol{\Sigma}_{B,u}, \boldsymbol{\Sigma}_{W,u}\}$ of the covariance matrices $\{\boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$. Given a collection of parameters θ and its estimate $\hat{\theta}$, RMSE is defined as

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{J}(\hat{\theta} - \theta)^\top(\hat{\theta} - \theta)}, \tag{6}$$

where J denotes the total number of scalar parameters collected in θ. For MSL, point estimates of ζ, $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$ 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 $\{\boldsymbol{\mu}_{1:N}, \boldsymbol{\beta}_{1:N,1:T}\}$, we use the sample mean $\zeta_0 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\mu}_n$ and the sample covariances $\boldsymbol{\Sigma}_{B,0} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\mu}_n - \zeta_0)(\boldsymbol{\mu}_n - \zeta_0)^\top$ and $\boldsymbol{\Sigma}_{W,0} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}(\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)(\boldsymbol{\beta}_{nt} - \boldsymbol{\mu}_n)^\top$ as true parameter values for ζ, $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$, respectively.

3.2.2 Predictive accuracy

To assess the predictive accuracy of the Bayesian methods, 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 *between* individuals. To that end, we generate a test set consisting of 100 observations from 100 new individuals along with each training sample. The realised choice and attributes of this sample are denoted by y_n^* and X_n^*. In the second scenario, we predict choice probabilities for new choice sets for individuals who already in the training sample and thus have a record of past choices, i.e. we predict *within* individuals. To that end, we generate another test set by creating an additional choice set for 100 individuals from the training sample. The realised choice and attributes of this sample are denoted by $y_n^\dagger$ and $X_n^\dagger$.

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

$$\text{BS} = \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\mathbf{1}\{y_n = j\} - \hat{P}_{nj} \right)^2, \tag{7}$$

where $\mathbf{1}\{y_n = j\}$ is an indicator, which equals one if the condition inside the braces is true and zero otherwise, and where $\hat{P}_{nj}$ 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. By contrast, 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 being that a higher value of P^{chosen} indicates superior predictive performance.

For mixed logit with unobserved inter- and intra-individual heterogeneity, the estimated predicted choice probabilities for the between-individuals prediction scenario are given by

$$\hat{P}(y_n^* | \mathbf{X}_n^*, \mathbf{y}) = \int \left(\int P(y_n^* | \mathbf{X}_n^*, \boldsymbol{\beta}) f(\boldsymbol{\beta} | \boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_W) d\boldsymbol{\beta} \right) f(\boldsymbol{\mu} | \hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\Sigma}}_B) d\boldsymbol{\mu}, \tag{8}$$

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

$$\hat{P}(y_n^\dagger | \mathbf{X}_n^\dagger, \mathbf{y}) = \int P(y_n^\dagger | \mathbf{X}_n^\dagger, \boldsymbol{\beta}) f(\boldsymbol{\beta} | \hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_W) d\boldsymbol{\beta}, \tag{9}$$

where $\hat{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\Sigma}}_W$ denote the posterior means of $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_W$, respectively. 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 (8) and (9).

3.3 Implementation details

We implement the MSL and MCMC estimators by writing our own Python code.[2]

For MSL, the numerical optimisations are carried out with the help of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Nocedal and Wright, 2006) contained in Python's SciPy library (Jones et al., 2001) and analytical gradients are supplied (see Appendix A.1 for details). We use 250 inter-individual simulation draws per decision-maker and 250 intra-individual simulation draws per observation. The simulation draws are generated via the Modified Latin Hypercube sampling (MLHS) approach (Hess et al., 2006). To assure that the covariance matrices maintain positive-definiteness, the optimisations are performed with respect to the Cholesky factors of the covariance matrices. For MSL, we also take advantage of Python's parallel processing capacities to improve the computational efficiency of the method. We process the likelihood computations in ten parallel batches, each of which corresponds to 25 inter-individual simulation draws.

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

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

In Table 1, we compare the predictive accuracy of the models that were estimated using Bayesian methods. For each experimental scenario S 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 between- and within-individuals 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 and the considered methods, we do not observe any significant differences in between-individuals predictive accuracy. As expected, standard logit without individual-specific parameters yields the same predictive accuracy in the between- and the within-individuals prediction scenarios. Due to the presence of individual-specific parameters, both mixed logit models improve the within-individuals predictive accuracy of standard logit by a significant margin. For instance, in scenario 1 for $T = 20$, MNL produces an average Brier score of 0.200, while mixed logit with only inter-individual heterogeneity and mixed logit with both inter- and intra-individual heterogeneity yield Brier scores of 0.152 and 0.149, respectively. Another insight is that the within-individuals predictive accuracy of mixed logit improves relative to standard logit, as more choice situations are included in the estimation. For example, in scenario 1, the Brier 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 heterogeneity does not offer significantly more accurate within-individuals predictions 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 affect the within-individuals prediction performance of considered mixed logit models. For example, for $T = 20$, the average Brier score for the within-individuals prediction scenario of mixed logit with unobserved inter- and intra-individual heterogeneity is 0.149 and 0.150 in both scenario 1 ($\alpha = 0.3$) and scenario 2 ($\alpha = 0.7$). We highlight that even in even in scenario 2, in which intra-individual taste variation accounts for 70% of the total variance in tastes, mixed logit with unobserved inter- and intra-individual heterogeneity does not outperform simple mixed logit with only inter-individual heterogeneity.

			Brier_{B}		Brier_{W}		$\text{P}_{\text{B}}^{\text{chosen}}$		$\text{P}_{W}^{\text{chosen}}$	
S	T	Method	Mean	SE [%]	Mean	SE [%]	Mean	SE [%]	Mean	SE [%]
1	10	MNL (MCMC)	0.202	0.200	0.201	0.229	0.394	0.311	0.395	0.372
		MXL-inter (MCMC)	0.199	0.222	0.165	0.318	0.401	0.336	0.541	0.606
		MXL-inter-intra (MCMC)	0.198	0.235	0.163	0.306	0.404	0.361	0.539	0.581
1	20	MNL (MCMC)	0.199	0.156	0.200	0.187	0.398	0.246	0.397	0.278
		MXL-inter (MCMC)	0.196	0.170	0.152	0.392	0.406	0.274	0.567	0.708
		MXL-inter-intra (MCMC)	0.196	0.181	0.149	0.392	0.409	0.296	0.570	0.708
2	10	MNL (MCMC)	0.202	0.206	0.201	0.226	0.394	0.339	0.395	0.375
		MXL-inter (MCMC)	0.199	0.220	0.167	0.293	0.401	0.351	0.536	0.542
		MXL-inter-intra (MCMC)	0.199	0.224	0.164	0.285	0.404	0.355	0.536	0.512
2	20	MNL (MCMC)	0.202	0.197	0.206	0.208	0.394	0.329	0.388	0.347
		MXL-inter (MCMC)	0.200	0.201	0.152	0.391	0.401	0.321	0.565	0.662
		MXL-inter-intra (MCMC)	0.199	0.209	0.150	0.380	0.404	0.327	0.566	0.636

Note: The reported values are averages and standard errors across 20 replications. S = experimental scenario. T = observations per individual. Brier = Brier score. P^{chosen} = average predicted probability of chosen alternative. B = between-individuals. W = within-individuals.

Table 1: Predictive accuracy on simulated data

Furthermore, Table 2 contrasts 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 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 are approximately tenfold with 3,423 seconds and 2,933 seconds, respectively.

			Time [s]	
S	T	Method	Mean	SE
1	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)	2932.7	86.9
1	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)	5003.8	122.2
2	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)	3085.6	162.7
2	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)	4372.6	156.0

Note: The reported values are averages and standard errors across 20 replications. S = experimental scenario. T = observations per individual.

Table 2: Estimation time on simulated data

Last, Table 3 compares the estimation accuracy of the MSL and MCMC estimators of mixed logit with unobserved inter- and intra-individual heterogeneity. Overall, we observe minor differences between the two estimators across the considered experimental scenarios. MSL is slightly more accurate than MCMC when the number of choice tasks per decision-maker is $T = 10$, while the converse seems to hold when $T = 20$.

			RMSE(ζ)		RMSE($\mathbf{\Sigma}_B$)		RMSE($\mathbf{\Sigma}_W$)	
S	T	Method	Mean	SE [%]	Mean	SE [%]	Mean	SE [%]
1	10	MXL-inter-intra (MCMC)	0.037	0.491	0.072	1.032	0.071	0.694
		MXL-inter-intra (MSL)	0.034	0.310	0.062	0.590	0.063	0.388
1	20	MXL-inter-intra (MCMC)	0.023	0.268	0.036	0.293	0.039	0.255
		MXL-inter-intra (MSL)	0.032	0.322	0.044	0.236	0.038	0.259
2	10	MXL-inter-intra (MCMC)	0.044	0.530	0.086	1.253	0.079	0.763
		MXL-inter-intra (MSL)	0.040	0.409	0.073	0.862	0.073	0.592
2	20	MXL-inter-intra (MCMC)	0.021	0.289	0.041	0.511	0.040	0.313
		MXL-inter-intra (MSL)	0.034	0.370	0.049	0.347	0.041	0.323

Note: The reported values are averages and standard errors across 20 replications. S = experimental scenario. T = observations per individual. RMSE = root mean square error.

Table 3: Estimation accuracy 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 and utility specification

Data for the empirical application are sourced from a stated preference survey on choices of holiday packages (Keane and Wasi, 2013, Louviere et al., 2008).[3] The data include observations from 683 respondents who each completed 32 choice tasks, which involved a choice of the best alternative of two unlabelled holiday packages characterised by 16 attributes. The final model specification includes nine attributes. Tastes with respect to five of these attributes are treated as fixed utility parameters, and taste with respect to the remaining four attributes are treated as random parameters. Table 4 provides a description of the considered attributes and shows which attributes pertain to fixed taste parameters parameters and which attributes pertain to random parameters.

Attribute	Levels
Attributes with fixed taste parameters	
Price	0 (\$999), 1 (\$1200)
Meal inclusion	0 (no), 1 (yes)
Distance from hotel to attractions	0 (200m), 1 (5km)
Local tours available	0 (no), 1 (yes)
Individual tour	0 (organised tour), 1 (individual)
Attributes with random taste parameters	
Overseas destination	0 (Australia), 1 (Overseas)
Length of stay	0 (7 days), 1 (12 days)
4-star accommodation	0 (2-star), 1 (4-star)
Beach or pool available	0 (no), 1 (yes)

Table 4: Attributes and levels of holiday package stated choice data

The data are randomly split into a training set and two test sets. The training set includes 20 choice tasks from each of 633 respondents. One test set is used to evaluate the between-individuals predictive ability of the considered models. It includes one choice task from each of the remaining 50 respondents. The second test set is used to evaluate the within-individuals predictive ability. It is formed by randomly selecting one of the remaining choice tasks from each of 200 respondents in the training sample. We create ten of such random splits and compare the performance of the different choice models across these splits.

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

$$U_{ntj} = \left(X_{ntj}^{\text{random}}\right)^\top \beta_{nt} + \left(X_{ntj}^{\text{fixed}}\right)^\top \gamma + \varepsilon_{ntj}. \tag{10}$$

[3] The data are publicly available in the online supplement of Keane and Wasi (2013).

Here, $\mathbf{X}_{ntj}^{\text{random}}$ is vector of attributes with individual- and observation-specific random taste parameters $\boldsymbol{\beta}_{nt}$, and $\mathbf{X}_{ntj}^{\text{fixed}}$ is a vector of attributes with fixed taste parameters $\boldsymbol{\gamma}$. ε_{ntj} is a stochastic disturbance with distribution $\text{Gumbel}(0, 1)$. The utilities of standard logit and mixed logit with only inter- and intra-individual heterogeneity are specified analogously.

4.2 Results

In Table 5, we compare the predictive accuracy of those discrete choice models which were estimated using Bayesian methods. 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 between- and the within-individuals prediction scenarios across ten random splits of the considered choice data. Overall, results are consistent with the results of the simulation evaluation. We do not observe any noteworthy differences in between-individuals predictive accuracy across methods. Both mixed logit models offer better within-individuals predictive accuracy than standard logit, but the more complex mixed logit model with unobserved inter- and intra-individual heterogeneity does not provide any benefits over the simpler mixed logit model with only inter-individual heterogeneity.

	Brier_B		Brier_W		P_B^{chosen}		P_W^{chosen}	
Method	Mean	SE [%]	Mean	SE [%]	Mean	SE [%]	Mean	SE [%]
MNL (MCMC)	0.221	0.843	0.214	0.395	0.562	0.868	0.571	0.371
MXL-inter (MCMC)	0.222	0.845	0.183	0.418	0.561	0.882	0.647	0.372
MXL-inter-intra (MCMC)	0.222	0.857	0.183	0.421	0.561	0.894	0.646	0.395

Note: The reported values are averages and standard errors across ten random splits. Brier = Brier score. P^{chosen} = average predicted probability of chosen alternative. B = between-individuals. W = within-individuals.

Table 5: Predictive accuracy on real data

Furthermore, Table 6 enumerates detailed estimation results for one the random splits of the stated choice data. The fixed taste parameters and the means of the random taste parameters of the three mixed logit models have the same signs. For mixed logit with unobserved inter- and intra-individual heterogeneity, the MSL and MCMC estimates 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 affords interesting behavioural insights into the sources of taste variation. We find evidence of substantial intra-individual taste variation. For example, MCMC suggests that $\frac{5.240}{6.149+5.240} = 46.0\%$ variation in tastes with respect to the attribute "overseas destination" are due to intra-individual heterogeneity. Similarly, MSL indicates that $\frac{2.943}{2.657+2.943} = 52.6\%$ of variation in tastes with respect to the attributes "4-star accommodation" can be ascribed to intra-individual heterogeneity.

	MNL		MXL-inter		MXL-inter-intra		MXL-inter-intra	
	(MCMC)		(MCMC)		(MCMC)		(MSL)	
Parameter	Mean	SD	Mean	SD	Mean	SD	Est.	SE
Fixed parameters								
Price	0.175	0.019	-0.445	0.035	-0.734	0.090	-0.732	0.077
Meal inclusion	0.370	0.019	0.560	0.029	0.972	0.136	0.949	0.122
Distance from hotel to attractions	0.666	0.018	-0.184	0.028	-0.311	0.059	-0.307	0.051
Local tours avail.	0.189	0.012	0.187	0.031	0.298	0.058	0.294	0.052
Individual tour	-0.315	0.024	0.215	0.030	0.350	0.060	0.350	0.049
Random parameters: Means								
Overseas destination	0.412	0.021	0.298	0.066	0.545	0.139	0.546	0.129
Length of stay	-0.145	0.020	0.569	0.045	1.005	0.157	0.947	0.140
4-star accommodation	0.144	0.026	0.913	0.050	1.534	0.223	1.460	0.197
Beach or pool avail.	0.177	0.026	0.234	0.023	0.363	0.058	0.362	0.059
Random parameters: Inter-respondent covariance								
Overseas destination vs. overseas destination			1.839	0.183	6.149	1.558	5.740	1.425
Length of stay vs. overseas destination			0.225	0.084	0.737	0.308	0.774	0.319
Length of stay vs. length of stay			0.430	0.073	1.408	0.397	1.205	0.351
4-star accommodation vs. overseas destination			0.084	0.087	0.247	0.300	0.344	0.281
4-star accommodation vs. length of stay			0.017	0.058	0.107	0.189	-0.096	0.173
4-star accommodation vs. 4-star accommodation			0.990	0.099	2.923	0.815	2.657	0.703
Beach or pool avail. vs. overseas destination			0.029	0.039	0.097	0.129	0.090	0.129
Beach or pool avail. vs. length of stay			-0.005	0.024	-0.035	0.074	-0.077	0.076
Beach or pool avail. vs. 4-star accommodation			-0.022	0.030	-0.093	0.102	-0.114	0.113
Beach or pool avail. vs. beach or pool avail.			0.096	0.020	0.260	0.086	0.218	0.070
Random parameters: Intra-respondent covariance								
Overseas destination vs. overseas destination					5.240	1.743	4.751	1.473
Length of stay vs. overseas destination					0.176	0.298	0.168	0.438
Length of stay vs. length of stay					0.191	0.196	0.114	0.190
4-star accommodation vs. overseas destination					0.278	0.543	0.401	0.505
4-star accommodation vs. length of stay					-0.227	0.195	-0.499	0.405
4-star accommodation vs. 4-star accommodation					3.335	1.603	2.943	2.586
Beach or pool avail. vs. overseas destination					0.119	0.217	0.379	0.319
Beach or pool avail. vs. length of stay					0.055	0.045	0.098	0.181
Beach or pool avail. vs. 4-star accommodation					-0.157	0.190	-0.154	0.774
Beach or pool avail. vs. beach or pool avail.					0.188	0.201	0.315	1.070

Note: For MCMC, the posterior mean and the posterior standard deviation are reported. For MSL, the point estimate and the asymptotic standard error are reported. For MSL, the standard errors of the covariance elements are obtained using a parametric bootstrap with 10,000 draws.

Table 6: Estimation results for one of the random splits of the real data

Finally, Table 7 gives the estimation times of the choice models across the ten random splits. Mixed logit with only inter-individual heterogeneity is substantially faster than mixed logit with unobserved inter- and intra-individual heterogeneity. MSL is slower than MCMC due to the presence of fixed utility parameters.

	Time [s]	
Method	Mean	SE
MNL (MCMC)	89.0	2.3
MXL-inter (MCMC)	312.5	6.0
MXL-inter-intra (MCMC)	2971.5	129.7
MXL-inter-intra (MSL)	7389.5	462.2

Note: The reported values are averages and standard errors across ten random splits.

Table 7: 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 within-individuals predictive 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 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 based on other consumers' preferences of while also exploiting interdependencies between products. Matrix factorisation provides a mapping of both consumers and products into a joint latent factor space and learns a sparse matrix of dimension # of consumers × # of products. Each cell of this matrix represents one consumer's preference that each consumer has for each product, which is a function of a sum of the product of a latent vector of alternative characteristics and a latent vector of consumer preferences for each of those 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 i) focus on prediction, instead of inference; ii) no underlying economic theory or lack of understanding about the relation between matrix factorisation and canonical models based on RUM theory; iii) inability to model time-varying choice sets and product-specific attributes. Economists and machine learning researchers came together recently to address 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 mobile location data. The main idea 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. The framework thus 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 such 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. Whereas this framework is already a good alternative to the mixed logit model with inter-and intra-heterogeneity, it can further be improved by taking inspiration from Athey et al. (2018) and incorporating the latent structure of matrix factorisation in 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 tasks per individual is large.

Besides, the estimation of mixed logit with unobserved inter- and intra-individual heterogeneity demands at least ten 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 review several promising alternative modelling approaches, which may offer superior prediction performance by flexibly capturing dependencies between decision-makers, alternatives and attributes.

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.

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics*, 89(1-2):57–78.

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.

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. (1998). Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transportation Research Part A: Policy and Practice*, 32(7):495–507.

Bhat, C. R. (2000). Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. *Transportation science*, 34(2):228–238.

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.

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*.

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.

Keane, M. and Wasi, N. (2013). Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6):1018–1045.

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.

Louviere, J. J., Islam, T., Wasi, N., Street, D., and Burgess, L. (2008). Designing discrete choice experiments: do optimal designs come at a price? *Journal of Consumer Research*, 35(2):360–375.

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.

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.

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., 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.

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 Model estimation

A.1 Maximum simulated likelihood

In maximum simulated likelihood (MSL) estimation, the mean vector ζ and the covariance matrices $\{\boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$ are treated as fixed, unknown parameters, whereas the individual- and case-specific parameters $\boldsymbol{\mu}_n$ and $\boldsymbol{\beta}_{nt}$ are treated as stochastic nuisance parameters. Point estimates of $\{\zeta, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$ 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 the covariance matrices in order to maintain positive-definiteness of $\{\boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W\}$.

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 integrating out the stochastic nuisance parameters:

$$LL(\theta) = \sum_{n=1}^{N} \ln\left(\int \prod_{t=1}^{T} \left(\int P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt}) f(\boldsymbol{\gamma}_{nt}|\boldsymbol{\Sigma}_W) d\boldsymbol{\gamma}_{nt}\right) f(\boldsymbol{\mu}_n|\zeta, \boldsymbol{\Sigma}_B) d\boldsymbol{\mu}_n\right), \quad (11)$$

where $\theta = \{\zeta, \mathbf{L}_B, \mathbf{L}_W\}$.

Since the integrals in expression 11 are not analytically tractable, we need to resort to simulation to approximate the log-likelihood. The simulated log-likelihood is given by

$$SLL(\theta) = \sum_{n=1}^{N} \ln\left(\frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R} P(y_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt,dr})\right)\right), \quad (12)$$

where $\boldsymbol{\beta}_{nt,dr} = \zeta + \mathbf{L}_B\boldsymbol{\xi}_{n,d} + \mathbf{L}_W\boldsymbol{\xi}_{nt,r}$. $\boldsymbol{\xi}_{n,d}$ and $\boldsymbol{\xi}_{nt,r}$ denote standard normal simulation draws. For each decision-maker, we take D draws for $\boldsymbol{\mu}_n$ and for each case, we take R draws for $\boldsymbol{\gamma}_{nt}$.

Point estimates $\hat{\theta}$ are then given by

$$\hat{\theta} = \arg\max_{\theta} SLL(\theta). \quad (13)$$

This optimisation problem can be solved with the help of quasi-Newton methods such as the limited-memory BFGS algorithm (see e.g. Nocedal and Wright, 2006). Quasi-Newton methods rely on the gradient of the objective function to find local optima. In principle, gradients can be approximated numerically. However, the numerical approximation of gradients is computationally expensive, as it involves many function evaluations. Computation times can be drastically reduced when analytical or simulated gradients are supplied to the optimiser. In the case of the mixed logit model with inter- and intra-individual heterogeneity, the two levels of integration impose a substantial computational burden, and thus efficient optimisation routines are critical for moderating estimation times.

In what follows, we derive expressions for the gradients of the mixed logit model with inter- and intra-individual heterogeneity. To the best of our knowledge, this is the first time these gradients are presented in the literature. First, we let ϑ_i denote one of the model parameters collected in $\boldsymbol{\theta}$. We have

$$\frac{\partial}{\partial\vartheta_i}\mathrm{SLL}(\boldsymbol{\Theta}) = \sum_{n=1}^{N} \frac{\frac{1}{D}\sum_{d=1}^{D}\frac{\partial}{\partial\varphi_i}\prod_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})\right)}{\frac{1}{D}\sum_{d=1}^{D}\prod_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})\right).} \tag{14}$$

To find the derivative in the numerator, we define

$$\psi_{nt,d}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R}P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr}) \tag{15}$$

with

$$\begin{aligned}\psi'_{nt,d}(\boldsymbol{\theta}) =& \frac{\partial\psi_{nt,dr}(\boldsymbol{\Theta})}{\partial\vartheta_i}\\ =& \frac{1}{R}\sum_{r=1}^{R}\left(P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})\frac{\partial V(\mathbf{X}_{ntj},\boldsymbol{\beta}_{nt,dr})}{\partial\vartheta_i}\right.\\ &\left.-\sum_{j'\in\mathcal{C}:j'\neq y_{nt}}\left(P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})P(j'|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})\frac{\partial V(\mathbf{X}_{ntj'},\boldsymbol{\beta}_{nt,dr})}{\partial\vartheta_i}\right)\right).\end{aligned} \tag{16}$$

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

$$V(\mathbf{X}_{ntj},\boldsymbol{\beta}_{nt,dr}) = \mathbf{X}_{ntj}^{\top}(\boldsymbol{\zeta}+\mathbf{L}_B\boldsymbol{\xi}_{n,d}+\mathbf{L}_W\boldsymbol{\xi}_{nt,r}), \tag{17}$$

we have

$$\frac{\partial V(\mathbf{X}_{ntj},\boldsymbol{\beta}_{nt,dr})}{\partial\boldsymbol{\zeta}} = \mathbf{X}_{ntj}, \tag{18}$$

$$\frac{\partial V(\mathbf{X}_{ntj},\boldsymbol{\beta}_{nt,dr})}{\partial\mathbf{L}_B} = \mathbf{X}_{ntj}\boldsymbol{\xi}_{n,d}^{\top}, \tag{19}$$

$$\frac{\partial V(\mathbf{X}_{ntj},\boldsymbol{\beta}_{nt,dr})}{\partial\mathbf{L}_W} = \mathbf{X}_{ntj}\boldsymbol{\xi}_{nt,r}^{\top}. \tag{20}$$

From the product rule of differentiation, it follows that

$$\frac{\partial}{\partial\vartheta_i}\prod_{t=1}^{T}\left(\frac{1}{R}\sum_{r=1}^{R}P(y_{nt}|\mathbf{X}_{nt},\boldsymbol{\beta}_{nt,dr})\right) = \left(\prod_{t=1}^{T}\psi_{nt,dr}(\boldsymbol{\theta})\right)\left(\sum_{t=1}^{T}\frac{\psi'_{nt,d}(\boldsymbol{\theta})}{\psi_{nt,d}(\boldsymbol{\theta})}\right). \tag{21}$$

A.2 Gibbs sampling

Under a fully Bayesian approach, the parameters $\boldsymbol{\zeta}$, $\boldsymbol{\Sigma}_B$, $\boldsymbol{\Sigma}_W$ are considered to be random, unknown quantities and are thus given priors. We use a normal prior $N(\boldsymbol{\xi}_0, \boldsymbol{\Xi}_0)$ for mean vector $\boldsymbol{\zeta}$ and Huang's half-t prior (Huang and Wand, 2013) for the covariance matrices $\boldsymbol{\Sigma}_B$ and $\boldsymbol{\Sigma}_W$. The latter is hierarchically defined: It consists of an inverse Wishart prior IW $(\nu + K - 1, 2\nu\boldsymbol{\Delta})$ with ν representing a known hyper-parameter and K denoting the number of underlying random parameters indexed by $k \in \{1,\ldots,K\}$. $\boldsymbol{\Delta} \equiv \mathrm{diag}(\boldsymbol{a})$

is a diagonal matrix with elements a_k distributed Gamma $\left(\frac{1}{2}, \frac{1}{A_k^2}\right)$. Akinc and Vandebroek (2018) show that Huang's half-t prior exhibits superior non-informativity properties compared to alternative prior specifications in the context of mixed logit with only inter-individual heterogeneity. In the subsequent applications, we set $\xi_0 = 0$, $\Xi_0 = 10^6 \mathbf{I}_K$, $\nu = 2$ and $A_k = 10^3 \; \forall k \in \{1, \dots, K\}$.

Stated succinctly, the generative process of mixed logit with unobserved inter- and intra-individual heterogeneity is as follows:

$$a_{B,k} | A_{B,k} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{A_{B,k}^2}\right), k = 1, \dots, K, \tag{22}$$

$$a_{W,k} | A_{W,k} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{A_{W,k}^2}\right), k = 1, \dots, K, \tag{23}$$

$$\boldsymbol{\Sigma}_B | \nu_B, \boldsymbol{a}_B \sim \text{IW}\left(\nu_B + K - 1, 2\nu_B \text{diag}(\boldsymbol{a}_B)\right), \quad \boldsymbol{a}_B = \begin{bmatrix} a_{B,1} & \dots & a_{B,K} \end{bmatrix}^\top \tag{24}$$

$$\boldsymbol{\Sigma}_W | \nu_W, \boldsymbol{a}_W \sim \text{IW}\left(\nu_W + K - 1, 2\nu_W \text{diag}(\boldsymbol{a}_W)\right), \quad \boldsymbol{a}_W = \begin{bmatrix} a_{W,1} & \dots & a_{W,K} \end{bmatrix}^\top \tag{25}$$

$$\zeta | \xi_0, \Xi_0 \sim \text{N}(\xi_0, \Xi_0) \tag{26}$$

$$\mu_n | \zeta, \boldsymbol{\Sigma}_B \sim \text{N}(\zeta, \boldsymbol{\Sigma}_B), n = 1, \dots, N, \tag{27}$$

$$\beta_{nt} | \mu_n, \boldsymbol{\Sigma}_W \sim \text{N}(\mu_n, \boldsymbol{\Sigma}_W), n = 1, \dots, N, t = 1, \dots, T, \tag{28}$$

$$y_{nt} | \beta_{nt}, X_{nt} \sim \text{Logit}(\beta_{nt}, X_{nt}), n = 1, \dots, N, \; t = 1, \dots, T, \tag{29}$$

where $\{\xi_0, \Xi_0, \nu_B, \nu_W, A_{B,1:K}, A_{W,1:K}\}$ are known hyper-parameters, and $\theta = \{\boldsymbol{a}_B, \boldsymbol{a}_W, \boldsymbol{\Sigma}_B, \boldsymbol{\Sigma}_W, \zeta, \mu_{1:N}, \beta_{1:N,1:T_n}\}$ is a collection of model parameters whose posterior distribution we wish to estimate.

The generative process given in expressions (22)–(29) implies the following joint distribution of the data and the model parameters:

$$\begin{aligned} P(y_{1:N}, \theta) = & \left(\prod_{n=1}^{N} \prod_{t=1}^{T_n} P(y_{nt} | \beta_{nt}, X_{nt}) P(\beta_{nt} | \mu_n, \boldsymbol{\Sigma}_W)\right) \left(\prod_{n=1}^{N} P(\mu_n | \zeta, \boldsymbol{\Sigma}_B)\right) \\ & P(\zeta | \xi_0, \Xi_0) P(\boldsymbol{\Sigma}_B | \omega_B, \mathbf{B}_B) \left(\prod_{k=1}^{K} P(a_{B,k} | s, r_{B,k})\right) \\ & P(\boldsymbol{\Sigma}_W | \omega_W, \mathbf{B}_W) \left(\prod_{k=1}^{K} P(a_{W,k} | s, r_{W,k})\right) \end{aligned} \tag{30}$$

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

$$P(\theta | y_{1:N}) = \frac{P(y_{1:N}, \theta)}{\int P(y_{1:N}, \theta) d\theta} \propto P(y_{1:N}, \theta). \tag{31}$$

Exact inference of this posterior distribution is not possible, because the model evidence $\int P(y_{1:N}, \theta) d\theta$ is not tractable. Hence, we resort to approximate inference methods. Becker et al. (2018) propose a Bayesian Markov chain Monte Carlo (MCMC) method in the form of a blocked Gibbs sampler for posterior inference in the described model. In what follows, we present the steps involved in one iteration of the sampler:

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

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

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

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

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

6. Update $\boldsymbol{\mu}_n$ for all $n \in \{1, \dots, N\}$ 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 $n \in \{1, \dots, N\}$ and $t \in \{1, \dots, T\}$:

   (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 $r = \frac{P(\mathbf{y}_{nt}|\mathbf{X}_{nt}, \tilde{\boldsymbol{\beta}}_{nt})\phi(\tilde{\boldsymbol{\beta}}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)}{P(\mathbf{y}_{nt}|\mathbf{X}_{nt}, \boldsymbol{\beta}_{nt})\phi(\boldsymbol{\beta}_{nt}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_W)}$.

   (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

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