Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

A metaheuristic-based extensive hypothesis testing for specification of latent class models

Prithvi Bhat Beeramoole[1], Ryan Kelly[2], Md. Mazharul Haque[1], Paul Scott[3], Alban Pinz[3], Alexander Paz[1,*]

[1]School of Civil & Environment Engineering, Queensland University of Technology, Australia

[2]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

[3]Department of Transport and Main Roads Queensland, Australia

Email for correspondence (presenting author): prithvi.beeramoole@.qut.edu.au ; prithvibhat.beeramole@hdr.qut.edu.au

1. Introduction

Methodological advances in discrete outcome modelling have largely focused on enabling representation of complex behaviors and improving forecasting accuracies. An important behavioral aspect often investigated is the presence of heterogeneity in the effects of contributory variables and how they vary in between alternative outcomes, across observations, and over discrete outcome events (Yuan et al., 2015). While multinomial-Logit models provide a fundamental basis for discrete outcome analysis under the random utility maximization framework, the basic structure comprises of shortcomings, including the ability to only capture effects that vary systematically with observed variables.

Over the years, research has focused on proposing improvements and flexibilities to the standard multinomial- Logit models to address some of the limitations. Among them, the Mixed-Logit models and Latent class models have by far been the most used to analyze heterogeneity in the effects. Mixed- Logit models use parametric distributions to accommodate effects that are heterogeneous (Vij and Krueger, 2017), offering additional insights regarding the disaggregate discrete processes. However, due to availability of several parametric distributions with different properties, the selection of an appropriate distribution has been recognized as an analyst-intensive task (Keane and Wasi, 2013, Vij and Krueger, 2017, Beeramoole et al., 2023, Paz et al., 2019). In contrasts, latent class models employ non- and semi- parametric distributions to relax some of the limitations of using parametric distributions, such as alleviating the need to prespecify the shape or functional form of the distribution (Vij and Krueger, 2017). However, the non-parametric distributions do not have a well-defined functional form. As a result, the specification problem translates from testing and selection of adequate parameter distributions in mixed-Logit models, to identification of optimal number of latent classes along with their corresponding membership and class-specific utilities.

Several such advanced specifications and associated mixing distributions are available today to address modelling limitations of the standard Logit and capture important and complex behavioral characteristics from the observed data, such as heterogeneity in tastes, nonlinearity, and correlation in the effects of variables on the discrete outcome. However, there is no consensus among researchers regarding the best approach to achieve this (Keane and Wasi, 2013). Further, prespecifying the specification structure without extensively testing all methods can potentially impose strong assumptions regarding behavior leading to biased or erroneous outcomes. To address this issue, there is need for a generalized framework to facilitate extensive testing of diverse hypotheses, while considering multiple methods to capture complex behavioral insights from the data.

In this study we propose an optimization-based framework to perform extensive hypothesis testing for estimation of discrete outcome models to adequately represent a combination of complex behaviors observed in an empirical setting. Previous studies (Paz et al., 2019, Ortelli et al., 2021, Beeramoole et al., 2023) have investigated the discrete outcome specification as an optimization problem. However, a detailed investigation on how the optimization-based framework can extract important behavioral information regarding the presence of heterogenous effects, while also examining the presence of latent classes, nonlinearities and correlation has not been conducted.

The proposed framework considers multiple specification types, including multinomial-, mixed-, and latent-class- Logit models, to generate unique hypotheses that simultaneously test potential explanatory variables, their functional forms, coefficients that capture heterogeneous preferences along with their mixing distributions, presence of latent segments with homogenous preferences within the observed data, optimal number of latent classes, presence of within-class heterogeneity in the effects and correlation. A metaheuristic-based solution algorithm is implemented to solve the proposed multi-objective optimization problem that evaluates the specifications based on both in- and out-of-sample fit. The proposed extensive hypothesis testing framework strategic and objective investigation of the data to capture important insights regarding behavior.

2. Methodology

2.1 Mathematical programming formulation

Latent class models typically constitute two main components, 1) a class-membership model, and 2) class-specific choice models that are conditional to class membership (Greene and Hensher, 2003). The class-specific models are defined using a mixed-Logit specification based on the utility maximization theory, as given by eqn. (1), which estimates the probability $P_{nj|q}$ of individual n choosing alternative j, conditional on n belonging to latent class q. The proposed specification allows simultaneous investigation of behaviors, including within-class nonlinearities, heterogeneity in preferences, and correlated effects. Since the class membership is unknown to the analyst, a prior membership probability $\bar{P}_{nq}$ of individual n belonging to q is estimated using eqn. (2). A multinomial-Logit specification is used to define class membership due to the discrete nature, with parameters of q^{th} class normalized to 0 to ensure model identification. In addition, the present formulation tests the estimation of class membership probabilities as both a function of a utility or as constants while ensuring that the membership probabilities for all classes sum to one. Further, when the total number of classes Q is equal to 1, the specification returns to a standard mixed- Logit model to test heterogeneity in the effects, and a multinomial-Logit model if only fixed coefficients are estimated.

To maintain interpretability, the latent class specification problem is formulated based on the approach used by Greene and Hensher (2013), wherein the vectors of individual characteristics $\mathbf{Z}_n$ and alternative attributes $\mathbf{X}_n$ are exclusively used to define class membership and class-specific models, respectively. However, depending on the context of the study, the latent class components can be appropriately defined using any of the variable vectors. Similarly, nonlinearity in class membership utilities can also be defined or tested.

The model coefficients $\boldsymbol{\beta}$ for $\mathbf{X}_n$ and $\boldsymbol{\theta}$ that are associated with $\mathbf{Z}_n$ are estimated using standard MLE (Train, 2003) and Expectation Maximization procedures (Train, 2008).

$$P_{nj|q} = \int \prod_{t=1}^{T} \frac{e^{\alpha_{jo}^{q}\beta_{jo}^{q}+\sum_{k=1}^{K}\alpha_{jk}^{q}\beta_{jk}^{q}{x_{njk}^{t}}^{(\lambda_k^q)}}}{\sum_j^J e^{\alpha_{jo}^{q}\beta_{jo}^{q}+\sum_{k=1}^{K}\alpha_{jk}^{q}\beta_{jk}^{q}{x_{njk}^{t}}^{(\lambda_k^q)}}} \mathbf{f}(\boldsymbol{\beta}_n^q)\mathbf{d}\boldsymbol{\beta}_n^q \tag{1}$$

$$\bar{P}_{nq} = \frac{e^{\ddot{\alpha}_0^q\theta_0^q+\sum_{m=1}^{M}\ddot{\alpha}_m^q\theta_m^q z_n}}{\sum_q^Q e^{\ddot{\alpha}_0^q\theta_0^q+\sum_{m=1}^{M}\ddot{\alpha}_m^q\theta_m^q z_n}} \tag{2}$$

In this study, the latent class specification problem is defined as a multi-objective non-linear mixed-integer combinatorial optimization problem, involving two conflicting objective functions – to minimize in-sample Bayesian Information Criteria (BIC) and minimize out-of-sample Mean Absolute Error (MAE), as given by eqns.(3) (3)and (4), respectively. Binary variables (5) are introduced in the specification to include or test specific features from the data in the membership and class-specific utilities to generate unique hypotheses.

Objective Functions:

$$\text{Min.}\,(\text{BIC}) = \delta\ln(N) - 2\left(\sum_{n}^{N}\ln\left[\sum_{q}^{Q}\frac{e^{\ddot{\alpha}_0^q\theta_0^q+\sum_{m=1}^{M}\ddot{\alpha}_m^q\theta_m^q z_n}}{\sum_q^Q e^{\ddot{\alpha}_0^q\theta_0^q+\sum_{m=1}^{M}\ddot{\alpha}_m^q\theta_m^q z_n}}\left(\int\prod_{t=1}^{T}\frac{e^{\alpha_{jo}^{q}\beta_{jo}^{q}+\sum_{k=1}^{K}\alpha_{jk}^{q}\beta_{jk}^{q}{x_{njk}^{t}}^{(\lambda_k^q)}}}{\sum_j^J e^{\alpha_{jo}^{q}\beta_{jo}^{q}+\sum_{k=1}^{K}\alpha_{jk}^{q}\beta_{jk}^{q}{x_{njk}^{t}}^{(\lambda_k^q)}}}\mathbf{f}(\boldsymbol{\beta}_n^q)\mathbf{d}\boldsymbol{\beta}_n^q\right)\right]\right) \tag{3}$$

$$\text{Min.}\,(\text{MAE}) = \frac{1}{J}\sum_{j}^{J}\left|\hat{s}_{v,j} - s_{v,j}\right| \tag{4}$$

Subject to:

$$\alpha_{jk}^q, \ddot{\alpha}_m^q, \omega_k^q, \phi_{k,p}^q, \hat{\alpha}_{jk}^q, \hat{\ddot{\alpha}}_m^q, \widehat{\omega}_k^q, \mu_k^q, \gamma_k^q, \hat{\phi}_{k,p}^q, \hat{Q} \in \{0,1\}\ \forall\ q, m, j\ , k, and\ p, p \neq k \tag{5}$$

Similarly, other constraints are imposed that test generic and alternative-specific effects of explanatory variables, their nonlinear transformations, and the correlation between their effects. In addition, constraints that allow pre-specification of part(s) of the membership and class-specific utilities are defined that enable analyst to test specific hypothesis or conduct a semi-guided specification search. pre-specifications ensure that the generated models align with the problem objectives and enable the consideration of important practical aspects beyond the statistics as often required in causal analyses.

2.2 Solution algorithm

In this study, Multi-objective Global-Best Harmony Search (MOGBHS) (Xiang et al., 2014) is adapted and integrated with the standard Maximum likelihood and Expectation Maximization parameter estimation methods to solve the proposed mathematical programming problem. Global-Best Harmony Search (GBHS) is a population-based metaheuristic, which combines the search strategies of Harmony Search and Particle Swarm Optimization to improve exploration of the solution space as well as intensify search near potential optimal solutions. The algorithmic steps designed to solve the proposed specification problem are described as follows.

```
Inputs: Dataframe containing $\mathbf{X}_n$, $\mathbf{Z}_n$, $y_{nj}^t\ \forall\ n \in \bar{N}, t \in \bar{T}, j \in \bar{J}$;
        Testing/validation dataframe, $v$
        Pre-specfciations: $\hat{Q}, \hat{\alpha}_{jk}^q, \hat{\ddot{\alpha}}_m^q, \hat{\omega}_k^q, \mu_k^q, \hat{\lambda}_k^q, \gamma_k^q, \hat{f}_k^q, \hat{\phi}_{k,p}^q$
Decision variables: $\alpha_{jk}^q, \ddot{\alpha}_m^q, \omega_k^q, \mathbf{f}^q, \mathbf{\Gamma}^q, \phi_{k,p}^q\ \forall\ m \in \bar{M}, j \in \bar{J}, k, p \in \bar{K}\ and\ p \neq k$;
        $Q,\ \boldsymbol{\beta}_n^q,\ \boldsymbol{\theta}^q, \lambda_k^q, \sigma_{k,p}^q\ \forall\ m \in \bar{M}, j \in \bar{J}, k, p \in \bar{K}\ and\ p \neq k$
Initialization
1.  Set initial values: $HMS,\ HMCR_{min}, HMCR_{max}, PAR_{min}, PAR_{max}, \rho, iter_{max}$
2.  Initialize an empty $MHM$
3.  Set $Q = 1, iter\ =\ 1$
Initialize harmony memory
4.  Initialize $HM_Q$ of size $HMS$ with random solutions; $HM_Q = [S_1, \ldots, S_{HMS}]$
5.  Evaluate objective functions $\forall\ S\ in\ HM_Q$
6.  Sort $HM_Q$ based on Fast non-dominant sorting and crowding distance
Improvise new harmony
7.  Repeat $iter_{max}$ times
    7.1. Harmony consideration
         Generate a random number ï ∈ $\{0,1\}$
         If ï ≤ $HMCR_{iter}$, then
               Generate a random number ş ∈ $\{1, \ldots, HMS\}$
               Randomly select features from solution $S_{ş}\ in\ HM_Q$ to create new solution $S_{iter}$
         else
               Generate a new random solution $S_{iter}$
    7.2. Pitch adjustment
         Generate a random number ï ∈ $\{0,1\}$
         If ï ≤ $PAR_{iter}$, then
               Perturb $S_{iter}$ by changing adding or removing a feature
         else
               No perturbation
    7.3. Evaluate objective functions
    7.4. $HM_Q \cup S_{iter}$
    7.5. Sort $HM_Q$ based on Fast non-dominant sorting and crowding distance
    7.6. Local Search
         If $iter\ \geq \rho \times (iter_{max})$, then
               Generate a random number ï ∈ $\{1, \ldots, len(PF_Q)\}$
               $S_{iter}\ =\ S_{ï}$
               Perturb $S_{iter}$ using pitch adjustment
               Go to step 7.7
         else
               Go to step 7.1
    7.7. Terminate improvise new harmony
         $iter\ =\ iter\ + 1$
         If $iter = iter_{max}$ , then
               Go to step 8
         else
               Go to step 7.6
Update Main harmony memory
8.  $MHM \cup HM_Q$
9.  Sort $MHM$ based on Fast non-dominant sorting and crowding distance
10. Keep only pareto front solutions in $MHM$;  $MHM = PF^{MHM}$
Termination
11. Estimate $\varrho_s$ using min-max normalization $\forall\ S$ in $PF_Q$
12. Estimate $\varrho_s^*$ using min-max normalization $\forall\ S$ in $MHM$
13. Find best solution $S_*^{PF_Q}$ in $PF_Q$, where $S_*^{PF_Q}$ is solution with min.($\varrho_s$)
    If $\varrho_s\ of\ S_*^{PF_Q} <\ max.(\varrho_s^*)\ \forall\ S$ in $MHM$, then
          Set $Q = Q\ + 1$
    else
    Return $HMH$
End
```

The specification search begins by setting the hyperparameters and initializing the harmony memory for class $Q = 1$. Initial solutions (M_1 ... M_{HMS}) are generated by randomly assigning values to binary variables (eqn. (6)) or using analyst pre-specifications. The solutions are then sorted from best to worst using Fast non-dominated sorting, proposed by Deb et al. (2002). An iterative process of 'improvising harmony' is then initiated, during which either some features from M_{iter} in memory are randomly selected and considered for improvisation or a new solution is generated. A pitch adjustment step follows, in which the decision variables in M_{iter} undergo perturbation based on a random number generator. Pitch adjustment allows testing of minute changes in the specification, such as inclusion or deletion of a variable, or change in random distribution for a specific variable's coefficient. The objective functions are evaluated, and the updated solution is included in HM followed by the non-dominant sorting of HM. A local search step is initiated as the iterations reach a pre-defined threshold ρ. During local search, only solutions within the Pareto-front (PF) are considered for improvisation. The process repeats for an increment in the number of latent classes. The specification search terminates when none of the solutions found during the current number of latent classes is observed in the PF solutions, or when the maximum number of latent classes is reached. The final PF is returned which contains the best set of non-dominant solutions.

3. Numerical experiment - transport mode choice preference using Swiss metro dataset.

3.1 Data description

The proposed MOGBHS was used to analyze transport mode choice behavior in Switzerland using a stated preference data collected by Bierlaire et al. (2001) in 1998. A detailed description of the dataset is provided by Antonini et al. (2007). Each respondent was presented with three transport mode alternatives (train, car, and Swiss metro), and nine hypothetical choice scenarios. Potential explanatory variables considered for the choice analysis included travel time (in minutes), travel cost (in CHF), headway for public transport modes (Train and Swiss metro), presence of luggage with traveler (no luggage, one, and more than one), seat configuration for Swiss metro (dummy variable indicating if the seats are arranged like airlines or not), dummy variable indicating if the traveler had an annual public transport ticket or not, traveler class (dummy variable to indicate first-class traveler), age, gender, income, and travel-cost bearer (self, employer, or both). For the experiments, 80% of the total observations (10,395) were used as the training dataset, while the remaining 20% were used to test out-of-sample prediction performance.

3.2 Results & analysis

Figure 1 presents the final Pareto front identified using MOGBHS with respect to in-sample BIC and out-of-sample MAE. Solutions found in the Pareto front were estimated with number of latent classes (Q) equal to two, which provided improvement in both BIC and MAE compared to those with only one class. The search terminated at $Q = 3$ as there was no significant improvement observed in BIC and MAE when compared to the Pareto Front solutions at $Q = 2$. Table 1 presents a relatively optimal solution selected from the elbow of the Pareto Front obtained using the MOGBHS.

The estimated specification identified two latent classes of travelers with distinct travel preferences. The class membership was defined using income levels, presence of luggage and a categorical variable that indicated if the travel cost was fully borne by travelers, or (partially or fully) subsidized by the employer. Latent class one is mostly likely to include travelers who prefer using public transport modes including Swiss metro and Train. The membership of travelers to this group is likely to reduce as their income levels and presence of luggage increase. In addition, they are less likely to be subsidized for travel costs. 35% of the observed sample are likely to belong to this group. The members of latent class one are not likely to prefer airplane like seat configurations and are sensitive to waiting time. The coefficient for headway was estimated with a normal distribution indicating significant heterogeneity in the effect of waiting time. Upto 85% of the observed sample associated a negative utility with headway. In contrasts, the Latent class two represented car dependents, who are likely to belong to high income groups and have access to travel costs subsidies. They are less likely to carry luggage and are sensitive to travel and waiting times. The random coefficients estimated for travel and waiting times indicate a significant variance in the associated preferences of travelers.

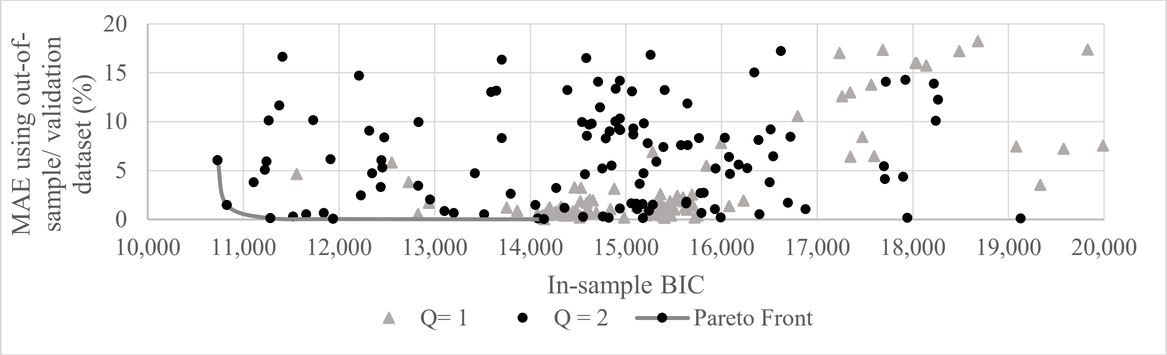

Figure 1: Pareto Front estimated using the MOGBHS algorithm

Table 1: Relatively optimal solution obtained from the Pareto Front for behavioral analysis for Swissmetro data

			Estimate	**t-ratio[1]**	**Estimate**	**t-ratio[1]**
		***f*[2]**	**Class1**		**Class2**	
Class membership Utilities						
Class-specific constant					0.44	1.7
Income					0.29	3.4***
Presence of luggage					-1.22	-8.3***
Travel cost subsidy					0.22	2.6***
Class-specific Utilities						
ASC Swissmetro			7.71	17.9***		
ASC Train			7.49	16.7***		
Travel time	mean				-0.03	-21.8***
	s.d.	*n*			0.04	-21.2***
Seats			-0.33	-2.7***		
Headway	mean		-0.05	-7.9***	-0.05	-18.9***
	s.d.	*n*	0.10	-12.6***	0.05	-17.0***
Log-Likelihood			**-5,370**			
AIC			**10,767**			
BIC			**10,830**			

*1.= weakly significant (p < 0.10, t > 1.645), ** = significant (p< 0.05, t>1.96), *** = strongly significant (p< 0.01, t>2.58)*
2.n = normal; u = uniform; t = triangular; ln = lognormal

The search approximately took 14 hours during which more than 300 unique specifications representing diverse combinations of behavior were tested. There is significant improvement in BIC from 14,837 to 10,830 during the search. However, the performance of the proposed MOGBHS was observed to drop when considering latent class specifications compared to when testing only for multinomial- and mixed-Logit specifications. This could be associated with the substantial increase in model complexity when latent class specifications are considered, which are prone to convergence issues.

Table 2 presents the non-dominant solutions found in the Pareto Front using the proposed extensive hypothesis approach. The BIC and MAE estimates show a trade-off between in-sample fit and out-of-sample prediction accuracy. While there is some similarity between the specifications based on the potential explanatory variables, the solutions vary in terms of complexity. These solutions can improve modelling efficiency by providing the analyst with objective start points to continue the model development based on the study context.

Table 2: Pareto Front solutions obtained from the search

	BIC	MAE	Specification type	Potential explanatory variables	Complexity
1.	10,731	6.03	latent class model with 2 classes	Cost, Headway, Seats, cost for traveler with annual public transport ticket Class membership: first class traveler, gender, income	Within-class heterogeneity with correlated parameters
2.	10,830	1.45	latent class model with 2 classes	Travel time, seats, and headway Class membership: income, presence of luggage, travel subsidy	Within-class heterogeneity in the effects of travel time and headway
3.	11,283	0.12	latent class model with 2 classes	Travel cost, cost for traveler with annual public transport ticket, alternative-specific constant Class membership: age, first class traveler, income, gender, travel subsidy	Within-class heterogeneity in the effects of cost-related variables
4.	11,941	0.04	latent class model with 2 classes	Headway, Seats, cost for traveler with annual public transport ticket	Within-class heterogeneity with correlated parameters
5.	14,144	0.02	mixed-Logit model	Travel time, cost, seats, age, income, gender, availability of luggage	Within-class heterogeneity in the effects of seat configuration

4. Conclusion

In this study, the generalized extensive hypothesis testing framework was proposed to include and test for advanced specifications such as those involving latent classes and within heterogeneous effects. The specification search considered multiple model performance measures including in-sample BIC and out-of-sample MAE to enable identification of superior specifications that best capture a combination of complex behaviors that are typically observed in real world datasets. The proposed method considers simultaneously multiple modelling decisions, including potential explanatory variables, their functional forms, the type of coefficients to be estimated, coefficients that capture heterogeneous preferences along with their mixing distributions, presence of latent segments with homogenous preferences within the observed data, optimal number of latent classes, presence of within-class heterogeneity in the effects and correlation. The experiment with Swissmetro dataset illustrated the performance of the proposed MOGBHS algorithm in providing important insights from the data regarding the contributory factors that affect latent class membership and the associated class-specific mode choice preferences. The estimated models revealed the presence of two latent classes of travelers with distinct preferences for transport modes. Planners and practitioners can potentially benefit from these insights as effective starting points to continue the hypothesis testing and subsequent model development.

References

ANTONINI, G., GIOIA, C. & FREJINGER, E. 2007. Swissmetro: description of the data.

BEERAMOOLE, P. B., ARTEAGA, C., PINZ, A., HAQUE, M. M. & PAZ, A. 2023. Extensive hypothesis testing for estimation of mixed-Logit models. *Journal of Choice Modelling,* 47**,** 100409.

BIERLAIRE, M., AXHAUSEN, K. & ABAY, G. 2001. The acceptance of modal innovation: The case of Swissmetro.

DEB, K., PRATAP, A., AGARWAL, S. & MEYARIVAN, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation,* 6**,** 182-197.

GREENE, W. H. & HENSHER, D. A. 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological,* 37**,** 681-698.

GREENE, W. H. & HENSHER, D. A. 2013. Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics,* 45**,** 1897-1902.

KEANE, M. & WASI, N. 2013. Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics,* 28**,** 1018-1045.

ORTELLI, N., HILLEL, T., PEREIRA, F. C., DE LAPPARENT, M. & BIERLAIRE, M. 2021. Assisted specification of discrete choice models. *Journal of Choice Modelling,* 39**,** 100285.

PAZ, A., ARTEAGA, C. & COBOS, C. 2019. Specification of mixed logit models assisted by an optimization framework. *Journal of choice modelling,* 30**,** 50-60.

TRAIN, K. 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press.

TRAIN, K. E. 2008. EM Algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling,* 1**,** 40-69.

VIJ, A. & KRUEGER, R. 2017. Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions. *Transportation Research Part B-Methodological,* 106**,** 76-101.

XIANG, W., AN, M., LI, Y., HE, R. & ZHANG, J. 2014. An improved global-best harmony search algorithm for faster optimization. *Expert Systems with Applications,* 41**,** 5788-5803.

YUAN, Y., YOU, W. & BOYLE, K. A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model. 2015.