Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

Can a metaheuristic be used to assist in discrete choice modelling?

Prithvi Bhat Beeramoole [1*], Alexander Paz [1], Md. Mazharul Haque[1], Alban Pinz[2]

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

[2] Manager (Economic Research and Analysis) at Department of Transport and Main Roads

Email for correspondence: prithvibhat.beeramole@hdr.qut.edu.au

1. Introduction

Discrete choice models have been an integral part of transport research including but not limited to road safety analysis, transport planning, land-use, and traffic operations. Even with the advent of advanced machine learning (ML) methods, discrete choice models are still widely used because of their ability to capture behavior and estimate causality. Over the years, extensions and capabilities have been added to better capture behavior, but each with their own strengths and limitations (Mannering et al., 2016).

The process of developing discrete choice specifications is highly involved as it requires the analyst to take critical modelling decisions, including (1) selection of variables to be tested during model specification; (2) identification of variable forms and transformations (e.g., linear or non-linear); (3) variables to be tested with fixed coefficients; (4) variables to be tested with random coefficients; (5) distributional assumptions for the random coefficients and error terms; (6) selection of methods to deal with potential correlation.

The decisions represent fundamental assumptions in the specification process, which significantly affect results and interpretation of underlying behavior. Most of the modelling decisions rely on experience and knowledge of the problem context. A common approach is to define restrictive specifications with some hypothesis and then iteratively modify until an acceptable model is obtained. Although the literature provides background to support these decisions, there is no certainty whether the estimated specification is the closest representation of the empirical truth. Furthermore, the decision burden on the analyst grows with model complexity. Therefore, such modelling approach can potentially introduce estimation errors and compromise model interpretability (Paz et al., 2019).

Most importantly, there is limited opportunity and resources for analysts to test large numbers of hypothesis and methods to address most modelling and data issues. Hence, restricted specifications with limited capabilities to address modelling challenges are frequently adopted. Further, the availability of highly dimensional datasets often makes the analysis more laborious and challenging. Considering that problem size grows substantially, an exhaustive search for a solution that address all data and modelling aspects is not feasible.

To assist analysts dealing with the above critical modelling decisions, an unbiased and efficient approach is required to discover important and meaningful information from the data. This study investigates whether a metaheuristic can simultaneously test critical modelling assumptions and assist in the development of flexible specifications to potentially seek insights beyond those that are often reported in the literature. The framework is expected to enable the investigation and estimation of various specifications in a large range of applications, including those involving multiple variables, and alternatives.

2. Methodology

2.1. Problem formulation

The observed utility associated with alternative $\boldsymbol{j}$ for individual $\boldsymbol{n}$ is given by $\boldsymbol{v_{nj}}$ in eqn. 1. Depending on the contribution to fit and intuitive behavioral meaning, the coefficients $\boldsymbol{\beta}$ for alternative attributes $\mathbf{X}_n$ can be estimated as generic, alternative-specific, fixed, random, or random-correlated coefficients. For outcome-independent characteristics, $\mathbf{Z_n}$ the corresponding coefficients θ are estimated as alternative-specific, with the base-outcome coefficients normalized to 0, ensuring that only $\boldsymbol{J-1}$ coefficients are estimated.

$$v_{nj} = \ddot{\boldsymbol{\alpha}}\boldsymbol{\theta}\mathbf{Z}_n + \boldsymbol{\alpha}\boldsymbol{\beta}\mathbf{X}_n \tag{1}$$

The model development process is considered as a non-linear mixed-integer combinatorial problem. The **objective function** is to minimize the Bayesian information criterion (BIC) given by eqn. 2.

$$\text{Min. BIC} = \delta \ln(N) - 2 * \sum_{n=1}^{N}\sum_{j=1}^{J} y_{nj}\left(\ln \int \frac{e^{\ddot{\boldsymbol{\alpha}}\boldsymbol{\theta}\mathbf{Z}_n + \boldsymbol{\alpha}\boldsymbol{\beta}\mathbf{X}_n}}{\sum_{j=1}^{J} e^{\ddot{\boldsymbol{\alpha}}\boldsymbol{\theta}\mathbf{Z}_n + \boldsymbol{\alpha}\boldsymbol{\beta}\mathbf{X}_n}} \mathbf{f}(\boldsymbol{\beta})\mathbf{d}\boldsymbol{\beta}\right) \tag{2}$$

subject to feature constraints that ensure the inclusion of specific features in the model. In addition, pre-specifications are imposed that allow analysts to include any knowledge a priori into the model development.

3. Solution Algorithm

A metaheuristic-based solution algorithm is proposed to solve the above mathematical programming problem using improved global-best harmony search (IGBHS) (Xiang et al., 2014). Hyperparameters are defined in the 'Initialization' step based on experimental trials and problem definition, which include: harmony memory size (HMS), minimum and maximum harmony memory consideration rate ($HMCR_{min}$ and $HMCR_{max}$); minimum and maximum pitch adjustment rate (PAR_{min} and PAR_{max}); maximum iterations ($iter_{max}$); threshold to initiate local search (ρ), and threshold to compare new solutions with solutions in HM (Δ).

A harmony memory (HM) of size HMS is initiated with randomly generated specification M. An opposition-based learning (OBL) algorithm is then implemented to initialize the opposite harmony memory. For each specification M in HM, an opposite model specification (OM) is generated using features that were not included in M to ensure extensive search. The two sets of random solutions significantly improve exploration. The two memories are then combined to generate a final harmony memory of size HMS after sorting the specifications based on the objective function. An 'Improvise harmony' step is initiated where the solutions in memory are perturbed based on dynamic values of $HMCR$ and PAR as given by Xiang et al. (2014). A new solution is created either by improving previous specifications in memory or by generating a new combination of decision variables. The pitch adjustment step follows wherein a new specification is fine-tuned by adding or removing some features. At each of these steps, the new specification is tested against the worst solution in memory, which is replaced if a better solution is found. The local search step is initiated towards the final stages when iterations reach a pre-defined threshold. A greedy-based strategy is deployed, wherein the best solution in memory is exploited to seek a better fit. For every change in the feature combination, the objective function is evaluated to check for an improvement in fit. If the new solution is unique

and better than any other solution in memory by a tolerance value Δ, the worst solution in HM is replaced. The tolerance value ensures a significant distinction between all solutions stored in memory. The search ends when the maximum number of iterations is reached, with the best-fit solutions in HM.

4. Experiments & Results

Mode choice preferences of travelers was analyzed using the proposed method for the dataset by Bierlaire et al. (2001) as proof-of-concept. The stated preference dataset was collected in Switzerland in 1998 to study the potential impact of a new transport mode – the Swiss metro. Each respondent was presented with nine hypothetical choice situations and asked to choose from three transport modes (train, car, and Swiss metro). 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). A detailed description of the dataset can be found in Antonini et al. (2007). Hyperparameters used for the experiments include: $HMS = 5$; $HMCR_{min} = 0.9$; $HMCR_{max} = 0.99$; $PAR_{min} = 0.8$; $PAR_{max} = 0.85$; $\mu = 80\%$; Δ = 15; and $iter_{max} = 300$.

Table 1 shows the Estimated Specification by the proposed solution algorithm along with the one estimated by Bierlaire et al. (2001). The Likelihood ratio test showed a significant improvement in fit by the specification estimated using the proposed algorithm (Chi-square score = 3,700; P-value < 0.00001 at 95% confidence interval). Variables such as travel time, travel cost and headway were identified as important and explanatory during the search, similar to the specification by Bierlaire et al. (2001). However, significant non-linearity in the effects of travel cost and headway was found by the Estimated Specification. In addition, socioeconomic characteristics, and trip-related attributes, including gender, availability of annual public-transport ticket, and travel-cost bearer were found to be significant factors influencing transport mode choices. While the alternative-specific constants in the specification by Bierlaire et al. (2001) suggest a higher preference for car, the Estimated Specification suggests a higher preference for train. More than 58% of the observed sample chose train as their preferred mode. The alternative-specific constants from the Estimated Specification possibly capture the unobserved utilities for train due to factors such as comfort. The socioeconomic characteristics included in the Estimated Specification reveal some interesting behavioral insights. For example, the large and positive coefficients for annual public-transport ticket indicate captive users, who perceive a high utility from public transport, which could possibly be due to subsidized travel cost.

The Estimated Specification suggests that male travelers associated a disutility with public transport modes, which possibly captures unobserved preferences of male travelers for factors including flexible departure times, convenience, and ability to perform chained trips. The Estimated specification also found that the observed sample were likely to choose Swiss metro if the travel cost was subsidized. The findings provide new venues for policy analysis, such as potential mode shift using incentives.

Significant heterogeneity in the effects of attributes, including travel cost, travel time and headway was captured during the search. Although the estimated individual-specific coefficients suggest an overall disutility associated with the attributes, the effect significantly varied across the observed sample. For example, the significant random coefficient for travel

cost suggests that some travelers had a greater dislike towards travel expenses, indicating the likely influence of other factors such as income levels and lifestyle preferences.

Table 1: Specifications by the proposed solution algorithm and the model from Bierlaire, et al. (2001)

		Specification by Bierlaire et al., (2001)		Estimated Specification by the proposed solution algorithm			
Number of respondents: 924 Number of observations: 8,316							
Parameter		Estimate	t-ratio[a]	Variable Transformation	Estimate	t-ratio[a]	*f*[b]
For Swiss metro							
Seat configuration		0.16	2*				
Annual public transport ticket		7.49	21.9***		56.61	19.6***	
Travel cost	mean	-0.001	-19.6***	Log	-4.41	-14.4***	
	s.d.				5.91	20.3***	*n*
Travel time	mean	-0.01	-24.3***	Square root	-0.53	-22.5***	
	s.d.				0.43	21.1***	*u*
Headway	mean	-0.01	-7.8***	Log	-0.72	-10.3***	
	s.d.				0.8	17.4***	*n*
For Car							
Alternative -specific constant		0.062	1.2		-2.17	-6.0***	
Male traveler					1.49	4.5***	
Luggage		-0.12	-2.5**				
Travel cost bearer					-0.34	-2.43**	
Travel cost	mean	-0.001	-19.6***	Log	-4.41	-14.4***	
	s.d.				5.91	20.3***	*n*
Travel time	mean	-0.01	-24.3***	Square root	-0.53	-22.5***	
	s.d.				0.43	21.1***	*u*
For Train							
Alternative -specific constant		-1.16	-10.4***		1.54	6.7***	
Annual public transport ticket		7.49	21.9***		56.61	19.6***	
Age		0.19	6.1***				
Male traveler					-1.06	-5.2***	
Travel cost bearer					-0.44	-4.0***	
Travel cost	mean	-0.001	-15.8***	Log	-4.41	-14.4***	
	s.d.				5.91	20.3***	*n*
Travel time	mean	-0.01	-15***	Square root	-0.53	-22.5***	
	s.d.				0.43	21.1***	*u*
Headway	mean	-0.007	-7.8***	Log			
	s.d.						*n*
Likelihood		**-6,565**				**-4,744**	
BIC		**13,211**				**9,575**	

a. * *= weakly significant (p < 0.10, t > 1.645),* ** *= significant (p< 0.05, t>1.96),* *** *= strongly significant (p< 0.01, t>2.58)*
b. n = normal; u = uniform

The final model was selected after a strategic search from a total of 1,156 specifications, which were estimated in twelve hours. Figure 1 shows convergence of the objective functions, BIC from 10,227 to 9,526, and LL from -5,052 to -4,715.

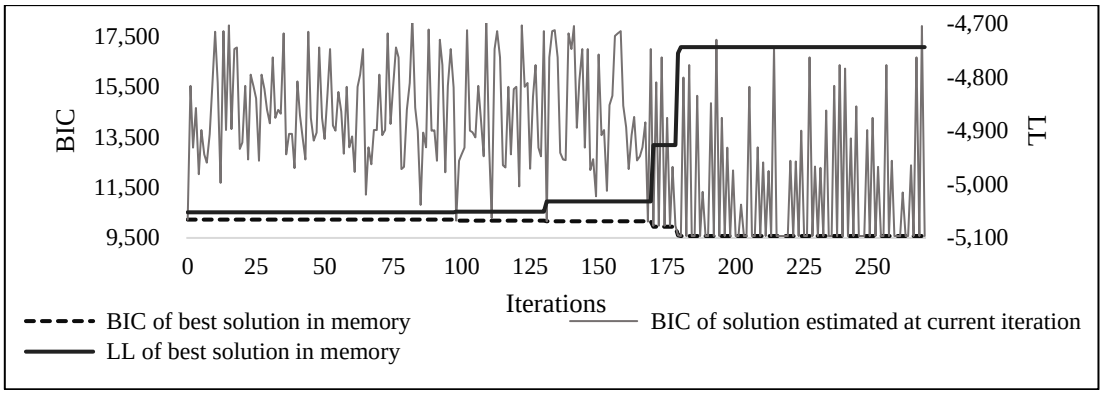

Figure 1: BIC vs Iterations for the experiment regarding the mode-choice preferences using Swiss metro dataset.

5. Conclusion

This study proposed a generalized framework to assist analysts in the estimation of discrete outcome models. This study addressed two important gaps in the literature while investigating the efficacy of the proposed solution algorithm. First, while previous studies focused on specific decisions during model development, the proposed formulation and associated solution algorithm enabled simultaneous testing of various modelling hypotheses including, the selection of potential explanatory variables, their functional forms, the distributional assumptions of coefficients, and correlations, thereby supporting estimation effort at a low cost. Second, the proposed formulation and solution algorithm provided flexibility to pre-specify certain modelling aspects to enable testing of specific hypotheses or ensure compliance with well-established theories from relevant fields, including economics and behavioral sciences. The experiment results illustrate the significance of the solution algorithm in discovering important influential or contributory factors, along with hidden patterns of nonlinearity, heterogeneity, and correlation, which can potentially be overlooked due to limited or biased search. A primary goal of any modelling is to capture as much information, insights, empirical truth, and underlying behavior as possible. However, a single "best" specification that explains all aspects of an empirical dataset may not exist. Hence, the proposed solution algorithm is useful as it can generate multiple acceptable solutions with varying properties and goodness-of-fit. The results validate that the solution algorithm can act as a decision-support tool by providing relevant starting points to the analyst for model development. However, regardless of the model development approach, analyst's knowledge and experience are necessary to guide the specification search in line with the study context. A potential extension to improve the applicability of the proposed solution algorithm could be the inclusion of discrete distributions into the framework to identify an optimum latent segmentation and estimate class-specific specifications. In addition, the framework could be extended to include advanced specifications such as those involving hybrid modelling methods.

6. Reference

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

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

MANNERING, F. L., SHANKAR, V. & BHAT, C. R. 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research,* 11**,** 1-16.

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

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