Exploration of model performances in the presence of heterogeneous preferences and random effects utilities awareness

Nikita Gusarov, Amirreza Talebijamalabad, Iragaël Joly

Working paper GAEL n° 12/2020

GAEL
Grenoble Applied Economic Laboratory
Consumption – Energy - Innovation

Exploration of model performances in the presence of heterogeneous preferences and random effects utilities awareness

Gusarov, Nikita
Talebijmalabad, Amirezza
Joly, Iragaël

October 2020

JEL codes: C25, C45, C52, C80, C90

https://gael.univ-grenoble-alpes.fr/accueil-gael

contact : agnes.vertier@inrae.fr

Exploration of model performances in the presence of heterogeneous preferences and random effects utilities

Nikita Gusarov[1] Amirreza Talebijamalabad[1] Iragaël Joly*

Abstract

This work is a cross-disciplinary study of econometrics and machine learning (ML) models applied to consumer choice preference modelling. To bridge the interdisciplinary gap, a simulation and theory-testing framework is proposed. It incorporates all essential steps from hypothetical setting generation to the comparison of various performance metrics. The flexibility of the framework in theory-testing and models comparison over economics and statistical indicators is illustrated based on the work of Michaud, Llerena and Joly (2012). Two datasets are generated using the predefined utility functions simulating the presence of homogeneous and heterogeneous individual preferences for alternatives' attributes. Then, three models issued from econometrics and ML disciplines are estimated and compared. The study demonstrates the proposed methodological approach's efficiency, successfully capturing the differences between the models issued from different fields given the homogeneous or heterogeneous consumer preferences.

Introduction

Consumer choices data are mainly modelled through classification tools from *Machine Learning* (ML) or econometric techniques. Economists and demand analysts deepen these analyses by studying consumers' willingness to pay (WTP). These economic measures are traditionally deduced from the assumed consumer behavioural theory underlying the estimated econometric model. In applications, the WTP are directly analysed or deduced from ML tools (for example, from recommendation system (Scholz et al. (2015)). These approaches of economic and behavioural indicators illustrate one of the differences between the two disciplines applying statistical learning. As described by Breiman and others (2001) and later by Athey and Imbens (2019): the ML focuses on the predictive qualities and econometrics attempts to decipher the underlying properties of the data. The hypothetico-deductive approach of econometrics allows the production of economic indicators under validity conditions of the model hypotheses, on which ML tools do not depend. This difference between approaches can then be viewed as a constraint or as an opportunity of the methods.

A specific focus of economists is the estimation of WTP of consumer for goods or goods' attributes. Multinomial regressions have been proposed in the literature to manage several behavioural assumptions, among which the heterogeneity across individuals by allowing taste parameters to vary in the population. Many choice experiments collect consumption choice data and analyse them with a mixed logit model which provides the advantage to consider such heterogeneity. Nevertheless many questions arise from the assumptions surrounding the introduction of the taste heterogeneity in the behavioural model (parametric or non parametric form, distribution choice of the parametric form, inter or intra-consumers heterogeneity (Hess and Rose (2008), Danaf, Atasoy, and Ben-Akiva (2020)), leading to many multinomial model competing specifications.

Switching focus to explanation of the findings clarifies why many of the advanced ML techniques rarely appear in economics publications. This is because of their believed lack of interpretability. Nevertheless, some pluri-disciplinary scientists make attempts to breach this wall between ML and econometrics: Athey and Imbens (2019), Mullainathan and Spiess (2017), Varian (2014). Their advances are mostly focused on the general interdisciplinary question, without entering into the application specific details.

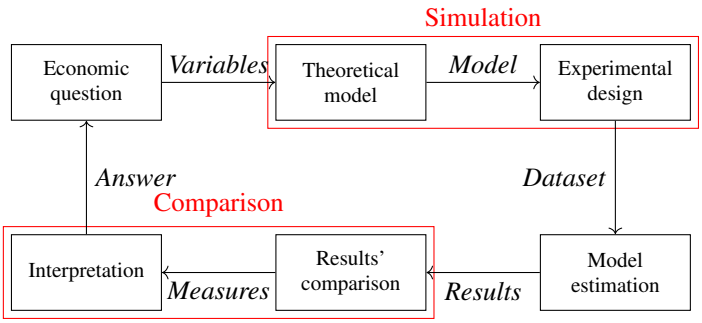
Objective of the paper is to evaluate and contrast performances of the two approaches, ML and econometrics, facing consumers preference heterogeneity. There have already been a multitude of studies comparing the performances of different econometric and ML models in various real world scenarios: the study of ML methods to model the car ownership demand estimation of Paredes et al. (2017); or the use of decision trees in microeconomics of Brathwaite, Vij, and Walker (2017). The performance of competing models are studied

*Univ. Grenoble Alpes, CNRS, INRAE, Grenoble INP, GAEL, 38000 Grenoble, France, email: iragael.joly@grenoble-inp.fr

according to several different criteria: (1) in terms of the quality of data adjustments; (2) in terms of predictive capacity; (3) in terms of the quality of the economic and behavioural indicators derived from estimates; and (4) according to their algorithmic efficiency and computational costs. However, to the best of our knowledge, there is no work providing a complete and comprehensive methodology for assessing and comparing model performances given the context of consumer preference studies.

The study proposes a theory-testing framework (Figure 1) exploring the performances of different econometric and ML models in presence of preference heterogeneity among individuals. More specifically, we assume the data are coming from choice experiment designed for value elicitation (Hensher, Rose, and Greene (2005), Louviere, Hensher, and Swait (2000)). In order to produce reliable results we construct two simulated datasets with homogeneous and heterogeneous preferences structures respectively. We start with the general methodological presentation of the undertaken procedure taken within the limits of the proposed theory testing framework. The second section on the work presents the simulation and estimation results, ending with a comparison of the performance metrics for the selected models. The last section concludes.

Figure 1: Proposed framework



Methodology

We aim to observe how some minor changes in theoretical decision model (specifically taste heterogeneity) may affect the results of value elicitation in choice experiment. We generate artificial datasets based on predefined behaviors and predetermined statistical properties for individual characteristics and alternatives' attributes. Such a set-up ensures that we know the exact data generation process and have all the control over the parameters and experimental design.

We use the work of Michaud, Llerena, and Joly (2012) to settle a realistic economic context of the empirical part of our work and to provide reasonable behavioural target values for the choice rules and tastes parameters. The work in reference investigates the impacts of the environmental attributes in the context of a consumer choice of non-alimentary agricultural goods. The Valentine's Day red rose is the good of the experiment. Subjects are faced with choice situation composed of two identical red roses by aspect with different specifications and an opt-out option. Product specifications are different among choice situations following random design technique. Among products attributes, *price* and two environmental aspects of roses' production are retained. The *eco-labelling* indicates the cultivation environment and conditions. The *carbon footprint* two-level factor measures the greenhouse gases emissions during the cultivation and transportation. The consumers are at the same time observed by four main socio-economic characteristics: *age*, *gender*, *income* and individual *habit* to acquire eco-labelled goods.

Artificial dataset

Generation of synthetic datasets is a common practice in many research areas. Such data is often generated to meet specific needs or certain conditions that may not be easily found in the real world data. The nature of the data varies according to the application field and includes text, graphs, social or weather data, among many others. The common process to create such synthetic datasets is to implement small scripts or programs, restricted to limited problems or to a specific application. In this work the simulation of the two datasets involves: (1) generation of an artificial population with characteristics issued from a set of predefined distributions; (2) creation of an experimental set-up based on a specific choice set ensemble; (3) simulation of the individual choices for given population and alternative sets using an arbitrary defined decision rule.

Following the reference paper, we consider the same four socio-economic characteristics. These four characteristics define the generated artificial population. For simplicity, we assume that these characteristics are not correlated. Sex and purchase habit are both binary variables generated separately with random draws from a Bernoulli distribution. To generate the class variables, age and income, we convert to the discrete-continuous multilevel scale draws from normal distribution defined by the mean and standard deviation parameters from the reference paper.

Stated choice (SC) experiments face sampled respondents with several different choice situations, each consisting of a comprehensive, and yet finite set of alternatives defined on a number of attribute levels. Based on this, respondents are asked to select their preferred alternatives given a specific hypothetical choice context. The experiment is designed in advance by assigning attribute levels to the attributes that define each of the alternatives which respondents are asked to consider (Rose and Bliemer 2008). In this research, we have implemented modified full factorial (FF) design following the ideas of the original paper, where the concern of reducing the number of choice situations' number was addressed. To make complete FF design taking into account the prices of the alternatives, we would have been faced with the nearly infinite number of distinct alternatives. To tackle this, we generate initial choice sets based on two binary variables using the FF design. We assume that individuals are presented with two unlabelled alternatives, roses *A* and *B*, as well as a no choice alternative (denoted *C*). The two attributes, eco-label and carbon footprint, have two levels which make four possible combinations for one alternative and 16 possible combinations in the case of multiple choice set-up (the no choice alternative has the levels fixed to zero). The prices are then randomly assigned to the predefined alternatives guiding the learning by adding potentially non-existant alternatives. Our simulated experimental design finally 'ask' the subjects to repeat 10 times their choices on new random designs in order to capture individual specific elements and achieve better statistical convergence.

Consumers' decisions are analysed with the discrete choice framework based on the utility maximisation assumption. This framework assumes that consumers associate each alternative in a choice set with a utility level and choose the option, which maximises this utility. The general estimation framework of the Random Utility Model (RUM) proposed by McFadden (1974) provides the opportunity to estimate the effects of product attributes (denoted as γ) and individual characteristics (β) and to compute WTP indicators. The deterministic part of utility function is given as follows[1]:

$$V_{ij} = \alpha_{i,Buy} + \beta_{Buy,Sex} Sex_i + \beta_{Buy,Age} Age_i + \beta_{Buy,Income} Income_i + \beta_{Buy,Habit} Habit_i + \\ + \gamma_{Price} Price_{ij} + \gamma_{i,Label} Label_{ij} + \gamma_{i,Carbon} Carbon_{ij} + \gamma_{i,LC} LC_{ij} \quad (1)$$

For different datasets the individuals are assumed to have homogeneous or heterogeneous preferences for the environmental attributes of alternatives. Each individual had his personal attitude to the eco-label and carbon footprint of the roses, determined by their awareness of the environmental issues.

In order to calculate utilities, we took parameters from the paper of reference (*a priori*). We started with calculating the relative deterministic utilities respectively for each individual and alternative, assuming that no choice option has zero utility for everyone. After adding some random noise, following the Gumble distribution parametrised with $(0, 1)$ we select the alternative with highest utility per each individual per each choice set. We took no-choice as reference alternative. This procedure is described in detail during obtained dataset presentation.

Modelling consumer choices

Adopted econometric models are multinomial logistic regression (MNL) and mixed multinomial logistic regression (MMNL), the later being of the possible generalisations of the former. The third model, a simplistic version of convolutional neural network (CNN), comes from the ML disciplines. Such models are rarely implemented by the economists in their studies since this family is usually perceived not to offer enough insight when it comes to the effects estimation. The ML techniques are usually viewed by economists as some black boxes, which do not provide any information about the underlying process. It is quite easy to comply with their position. Although the most advanced techniques perform better in terms of predictive power, they rarely offer any insight into the modelling process. The chosen CNN is adjusted to answer the economic question through modelling of the relative deterministic utility functions.

[1] Where $LC = Label \times Carbon$.

The two econometric models are perfectly adapted to model one of the two generated datasets respectively[2]. The MNL model should yield the best performance results on a dataset assuming fixed effects, while its counterpart, MMNL, should be the most performant in the presence of random effects in the utility function. The MNL model assumes that the decision makers view the available alternatives to be independent and that attribute impacts are fixed for the whole population across all alternatives (McFadden (1974)). This assumption is relaxed in the MMNL, where coefficients (or some of them) vary for each individual (Agresti 2013). The logistic regression models are derived from GLM specifications (Agresti 2007). This class of models relies on the hypothesis, that each individual maximises his perceived utility over a closed set of alternatives. His utility is assumed to be determined by a fixed and a random parts. The probability structure incorporates the theoretical assumptions of the finite choice set, the uniqueness of the chosen alternative and the idea of utility maximisation. Many of the existing applied econometrics papers use the most simple MNL model, which may lead to erroneous results and conclusions in the presence of random taste coefficients in the utility.

The model issued from the ML field focuses on more advanced and atypical modelling techniques. The neural network (NN) models can be viewed as an even wider generalisations of the generalised additive models (GAM), and are capable to imitate more simple models similar to MNL. The resulting CNN comprises two layers: (1) convolutional layer and (2) *softmax* transformation layer. The convolutional layer transforms the linear combination of individuals characteristics and alternatives' attributes into the relative deterministic utilities. Then, the utilities are passed to the *softmax* layer with fixed weights to derive the resulting choice probabilities. This choice was made since the seemingly identical models by their structure may produce different results depending on the implemented estimation techniques. The NN's offer us a great number of different algorithms which are more advanced than the algorithms traditionally implemented in econometrics, which make us wonder whether the changes in the estimation algorithm will allow us to achieve better results. In this study we use *Adam* algorithm (Kingma and Ba 2014) for CNN estimation, which is parametrized according to *Keras*[3] standards, with increased learning rate (fixed at $1e-1$).

Performance measures

In the first place we are interested by the overall goodness in estimation of the utility function components. In this task we should compare the obtained estimates with the target values we have settled into the utility functions. The best model should produce the mean estimates, which are equal to the targets, with the minimal variance possible.

Secondly, we are attracted by the WTP for roses and the premiums associated with particular alternative specific attributes. These were the only target metrics present in the article of Michaud, Llerena, and Joly (2012). The WTP could be read as the value the consumers are willing to pay for a rose. At the same time, the premiums may be translated as how much consumers are ready to pay for a unit change of a given attribute of the product. Both the WTP for a product and the premiums can be computed as the marginal rates of substitution between the quantity expressed by the attributes and the price (Louviere, Hensher, and Swait 2000). These theoretical values could be easily derived for all the three explored models, calculated as ratios of the corresponding coefficient (or weights). They will allow us to compare how close the derived values are to the theoretical input values, which were defined on the dataset generation step.

Thirdly, it is important to assess the overall goodness of fit over the whole dataset for the selected models. To address this issue, the best suited measure is the *accuracy*, describing the part of correctly classified instances in a given set and is by its nature a complement to the empirical error-rate measure (Japkowicz and Shah 2011). Doing so, we will be able to observe the ratio of the overall correctly modelled choices. We may as well implement the Kullback–Leibler Divergence (KL or KLD) estimator for overall goodness of fit. This will allow us to quantify the difference between the estimated posterior distribution and the true underlying distribution of the choices.

Finally, we observe the performance of these different models in terms of computational efficiency in resources consumption. For this task we will observe the computation times for given models. This measure is one of the most complex, because it accounts at the same time for different models, different estimation algorithms, different numerical implementations in the statistical software and different PC configurations. It is valid in this particular case, because all models were estimated using the same hardware and software set-up.

[2] For model estimation we use *mlogit* package, version 1.1-0 from CRAN

[3] Version 2.3.0.0 from CRAN

Table 1: The assumed relative utility function parameters

(a) Mean effects

	Effects
	Means
Characteristics (β)	
Sex	1.420
Age	0.009
Income	0.057
Habit	1.027
Attributes (γ)	
Price	-1.631
Buy	2.285
Label	2.824
Carbon	6.665
LC	-2.785

(b) Variance-covariance

	Effects	
	Fixed	Random
Variance		
Buy	0	3.202
Label	0	2.654
Carbon	0	3.535
LC	0	2.711
Covariance		
Buy:Label	0	-0.54
Buy:Carbon	0	-4.39
Buy:LC	0	6.17
Label:Carbon	0	8.77
Label:LC	0	-2.33
Carbon:LC	0	-4.82

Results

We present the obtained results in several steps. First of all a discussion on the simulated datasets is provided. Then we present the estimation results and present to the reader the goodness of relative deterministic utility function coefficients estimates for the different models. Finally, we provide an extensive discussion of the performance results.

Data

Each artificial datasets regroups 1000 artificial individuals, each of them faced with 16 different choices 10 times with random prices allocation (160 choice situations in total), hence, 160000 observations per dataset. In both situations the utility functions are determined as in paper: we use the exact means for the coefficients estimates, assuming they are correct (Table 1a). The variance-covariance matrix for RUM individual coefficients is supposed to be a matrix of zeros for the homogeneous preferences case and to be as in the reference paper for the heterogeneous preferences dataset (Table 1b). These coefficients are then randomly assigned within population with draws from a multivariate normal distribution.

It is interesting to explore the statistical properties of the two resulting artificial datasets and the original one, gathered by Michaud, Llerena, and Joly (2012)[4]. ANOVA and χ^2 tests (table 2) show no significant means difference between the simulated datasets and the original one, except for the *Income* variable. This is explained by the implemented dataset generation procedure based on transformation of draws from symmetric normal distribution. The distributions of $Carbon$ footprint and Eco-$Label$ attributes follow the ones inside the original dataset, while the distribution of price differs (table 3).
This particular divergence, may be explained by the procedure implemented to assign prices to the alternatives inside choice sets, because the random generator algorithms differ across statistical programs and potentially the procedures implemented in different softwares[5] are not identical.

Differences in the $Choice$ proportions appears interestingly. There is an important work in comparing the statistics for different classes in our sample to ensure that they are not biased in favour of label "A" or label "B," as in this case, the estimates are prone to be biased. For the artificial dataset the ratio of choices per "Buy" alternative is higher than 40% and reaches 47.3% for the fixed effect utility. At the same time, for the random effects specification, the numbers are lower, reaching only 42% in mean for two classes. This particular observation is rather interesting as it demonstrates how the heterogeneous tastes for alternatives' characteristics affect the consumer decisions.

[4]To save some space, these summary statistics are available upon request
[5]In this work the *R version 3.5.2 (2018-12-20) – "Eggshell Igloo"* was used.

Table 2: Individuals' descriptive statistics by dataset

	Fixed Effects	Random Effects	*Target*	p value
Sex				0.851
Mean	0.506	0.515	0.490	
SD	(0.500)	(0.500)	(0.502)	
Range	0 - 1	0 - 1	0 - 1	
Habit				0.182
Mean	0.683	0.657	0.604	
SD	(0.466)	(0.475)	(0.492)	
Range	0 - 1	0 - 1	0 - 1	
Income				< 0.001
Mean	2.750	2.671	2.147	
SD	(1.476)	(1.438)	(1.222)	
Range	1 - 6	1 - 6	1 - 6	
Age				0.255
Mean	41.862	42.161	39.755	
SD	(13.685)	(13.820)	(18.895)	
Range	18 - 84	18 - 84	18 - 85	

Table 3: Alternatives' descriptive statistics by dataset

	Fixed Effects	Random Effects	*Target*	p value
Price				0.002
Mean	2.936	2.936	3.005	
SD	(0.958)	(0.958)	(0.887)	
Range	1.5 - 4.5	1.5 - 4.5	1.5 - 4.5	
Carbon				0.999
Mean	0.5	0.5	0.5	
SD	(0.5)	(0.5)	(0.5)	
Range	0 - 1	0 - 1	0 - 1	
Label				0.999
Mean	0.5	0.5	0.5	
SD	(0.5)	(0.5)	(0.5)	
Range	0 - 1	0 - 1	0 - 1	

Estimation results

The comparison of the utility function coefficient estimates obtained by the different models over different datasets can be done in several steps. First of all, we are interested in the observed mean effects over the datasets, because the possibility to correctly identify the means for the coefficients is of outmost importance for the analysis, regardless of the assumption about homogeneity or heterogeneity of these effects. Then we explore the additional dimension provided by the MMNL estimates, which comprises the estimates for the variance-covariance matrix of the correlated random effects. Finally, we will give some comments on the CNN model estimates.

In the case of homogeneous preferences structure the MNL model obtains the exact estimates with fast a convergence rate and relative simplicity of the problem (table 4). The estimates obtained with the MMNL model for the fixed effects dataset demonstrate quasi-identical estimates to the MNL model. The only disadvantage of the MMNL models misspecification in this case resides in the significantly increased estimation time, which requires significantly more iterations in order to estimate correctly the variance-covariance matrix elements and, consequently, the estimation complexity.

In the case of heterogeneous preferences as estimates are significantly biased for the MNL model (table 4). The MNL model tends to significantly underestimate the effects of all of the characteristics and attributes for the choice situation. This can potentially lead to a significant bias in case we were using incorrect model specification during a field experiment data exploration. The estimates obtained with the MMNL model are slightly biased as well in this case.

Table 4: Estimation results

	Fixed effects			*Random effects*			*Target*
	MNL	MMNL	CNN	MNL	MMNL	CNN	
Characteristics							
Sex	1.401***	1.400***	1.369	0.712***	1.297***	0.719	1.420
	(0.031)	(0.031)		(0.016)	(0.024)		
Age	0.009***	0.009***	0.010	0.007***	0.010***	0.005	0.009
	(0.001)	(0.001)		(0.001)	(0.001)		
Salary	0.048***	0.048***	0.060	0.066***	0.120***	0.062	0.057
	(0.010)	(0.010)		(0.005)	(0.008)		
Habit	1.070***	1.071***	1.056	0.361***	0.641***	0.343	1.027
	(0.030)	(0.030)		(0.016)	(0.024)		
Attributes							
Price	-1.626***	-1.628***	-1.618	-0.886***	-1.586***	-0.886	-1.631
	(0.010)	(0.010)		(0.006)	(0.010)		
Buy	2.311***	2.313***	2.228	0.662***	2.180***	0.665	2.285
	(0.065)	(0.066)		(0.036)	(0.054)		
Label	2.815***	2.817***	2.810	1.279***	1.922***	1.277	2.824
	(0.022)	(0.022)		(0.015)	(0.023)		
Carbon	6.654***	6.662***	6.634	3.259***	5.430***	3.250	6.665
	(0.032)	(0.033)		(0.016)	(0.030)		
LC	-2.781***	-2.782***	-2.765	-1.546***	-2.663***	-1.558	-2.785
	(0.028)	(0.028)		(0.019)	(0.030)		
Note:							*p<0.1; **p<0.05; ***p<0.01

Even as the estimates of the means obtained with MMNL in the presence of the random effects are close to the theoretical ones, the estimates of the variance-covariance matrix elements are rather close, but not perfectly measured. This situation demonstrates the existing trade-off between the need to correctly specify the model from the start and the potential computation inconveniences in the case of implementation of a more complex model under uncertainty. In other words, there is always a choice either to simply use more complex model, which requires more data, calculation time and resources, or to perform an extensive theoretical study beforehand in order to correctly specify and delimit the model from the start.

Our CNN model is identical in structure to the MNL model, estimated with *Adam* algorithm. Because of the nature of the constructed CNN model, the obtained estimates in the presence of fixed effects are technically

Table 5: General performance measures

	MNL	MMNL	CNN
Accuracy			
FE	0.863	0.863	0.723
RE	0.725	0.863	0.721
KL			
FE	0.623	0.623	0.328
RE	0.349	0.625	0.317
Time			
FE	20.910	452.414	17.433
RE	18.722	2066.934	16.806
Note:	FE - fixed; RE - random effects		

identical to the estimates obtained with the MNL model. These results demonstrate the flexibility of the NN models and the hypothetical possibility to implement them in place of traditional econometric models with only inconvenience being the relative complexity to obtain the variances for the weights estimates, as no method known to us allows this, or to estimate variances through a cross-validated training of the NN. In the presence of random effects, the proposed CNN algorithm is, identically to MNL model, unable to correctly identify parameters and consequently derive the true means for the underlying coefficients of the relative utility function in the presence of heterogeneous preferences among individuals.

Performance comparison

Performance in terms of utility function estimation was presented in the previous section. Three complementary performance metrics are described: (1) the overall fit quality, (2) the computational efficiency and (3) the economic indicator precision.

First of all we focus our attention on the general performance metrics, describing how well the estimated models fit the predicted outcomes over an original dataset. We can observe the values of accuracy and KL divergence, describing overall performance of a given model, in Table 5. The table gathers the metrics' values for all the estimated models over both datasets. We observe quite natural situation: the best model in terms of overall performance is the model which is based on the choice rule used in the data generation step. The MNL and MMNL models perform equally well on the fixed effects dataset. This fact supports our initial hypothesis that an implementation of a more complex model is preferred when the real effects are unknown to the researcher. CNN model did not outperform the MNL. This observation may be explained by the data-generation set-up, where the generative algorithm favoured the MNL model with Gumbel error term rather than more general NN framework.

Table 5 presents the resources efficiency indicator: CPU time spent for execution by the system on behalf of the calling process. The more advanced *Adam* algorithm implementation with *Keras* easily outperforms the algorithms available in the *mlogit* package, although this boost in efficiency goes at the cost of lower overall performance and goodness of fit. At the same time, the MMNL implementation is far less efficient and takes 128 times more time, than CNN model. This situation clearly illustrates us how the precision and flexibility come at higher costs.

Finally we focus on the case specific metrics, WTP and premiums estimates present in the Table 6, that the consumers are eager to pay for particular environmental attributes. Comparing the estimates with the input values, we notice that the variances of the WTP and Premiums estimates (presented in brackets), estimated over a fixed effects dataset, do not potentially affect the conclusion one can derive from the results. We may conclude that given sufficiently large dataset the implementation of a more complex model (MMNL in this particular case) is preferable, because it will allow to control for unknown parameters without adding a risk of obtaining biased results. The simpler models, should be preferred in a more restricted context. They empower us to obtain valid results only in the case of correct theoretical assumptions, biasing the estimates in other conditions. Consequently, in the presence of uncertainty about the presence of heterogeneity in the customer choice modelling questions there is a strong interest to implement a more complex model, readjusting it afterwards if needed.

Table 6: Performance in terms of WTP and premiums

	Fixed effects			*Random effects*			*Target*
	MNL	MMNL	CNN	MNL	MMNL	CNN	
WTP							
Mean	1.421	1.416	1.377	0.747	1.360	0.751	1.401
SD		(0.058)			(1.887)		(1.973)
Label							
Mean	1.731	1.732	1.737	1.445	1.243	1.442	1.731
SD		(0.019)			(1.667)		(1.611)
Carbon							
Mean	4.091	4.097	4.101	3.679	3.467	3.669	4.086
SD		(0.103)			(2.323)		(2.134)
LC							
Mean	4.112	4.116	4.129	3.378	3.036	3.352	4.110
SD		(0.098)			(3.240)		(3.379)

Conclusion

In this work we have introduced the reader to the problematic of the different modelling paradigms in application to the consumer choice studies. By means of an experimental theory-testing framework we demonstrate the complexity of the model performance evaluation problematic, showing the eventual bottlenecks and the questions to be answered on all the levels of data exploration procedure. The correct specification of the theoretical assumptions, the dataset generation, the model choice as well as the performance measure choice were studied.

Given the experimental design and selected parameters, the MMNL model proves itself to be preferable. The ability to correctly estimate the target effects in presence of preference structure uncertainty is of great value in the field experiments. The CNN model illustrates the possibility for economists to implement the advanced ML techniques to treat economic questions.

One limitation of this work concerns the external validity of the observations. Arbitrary choices made in the study limit our conclusions to this specific case, and require more extensive experimentations to produce more general conclusions. Metaparameters of the framework will allow to specify sample size and compared tools. The presented results are conditionned with 1) the large sample size leading to highly significant estimates of MNL and MMNL ad 2) the CNN design aiming to reproduce the MNL, including its limitations. This work demonstrates only a fraction of the full potential of the theory-testing framework. Many extensions and generalisations should be performed before it could be used at scale. For example, it is particularly interesting to introduce an extension which will provide the possibility to explore and compare how different behavioural theories affect the estimation results. The framework could be complemented with a methodological tool-set for hypothesis testing using the advantages of a controlled experiment data collection as well.

Bibliography

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis, Second Edition*.

———. 2013. *Categorical Data Analysis, Third Edition*.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Brathwaite, Timothy, Akshay Vij, and Joan L Walker. 2017. "Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice." *arXiv Preprint arXiv:1711.04826*.

Breiman, Leo, and others. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Danaf, Mazen, Bilge Atasoy, and Moshe Ben-Akiva. 2020. "Logit Mixture with Inter and Intra-Consumer Heterogeneity and Flexible Mixing Distributions." *Journal of Choice Modelling* 35: 100188. https://doi.org/https://doi.org/10.1016/j.jocm.2019.100188.

Hensher, David A., John M. Rose, and William H. Greene. 2005. *Applied Choice Analysis: A Primer*. Cambridge University Press. https://doi.org/10.1017/CBO9780511610356.

Hess, Stephane, and John M. Rose. 2008. "Asymmetrical Preference Formation in Willingness to Pay Estimates in Discrete Choice Models." *Transportation Research Part E*, 847–63.

Japkowicz, Nathalie, and Mohak Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803.

Kingma, Diederik P, and Jimmy Ba. 2014. "Adam: A Method for Stochastic Optimization." *arXiv Preprint arXiv:1412.6980*.

Louviere, Jordan J, David A Hensher, and Joffre D Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge university press.

McFadden, Daniel. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4): 303–28. https://doi.org/https://doi.org/10.1016/0047-2727(74)90003-6.

———. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4): 303–28. https://doi.org/https://doi.org/10.1016/0047-2727(74)90003-6.

———. 2001. "Economic Choices." *The American Economic Review* 91 (3): 351–78. http://www.jstor.org/stable/2677869.

———. 2001. "Economic Choices." *The American Economic Review* 91 (3): 351–78. http://www.jstor.org/stable/2677869.

Michaud, Celine, Daniel Llerena, and Iragael Joly. 2012. "Willingness to pay for environmental attributes of non-food agricultural products: a real choice experiment." *European Review of Agricultural Economics* 40 (2): 313–29. https://doi.org/10.1093/erae/jbs025.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Paredes, Miguel, Erik Hemberg, Una-May O'Reilly, and Chris Zegras. 2017. "Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?" In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 780–85. IEEE.

Rose, John M, and Michiel CJ Bliemer. 2008. "Stated Preference Experimental Design Strategies." *Handbook of Transport Modelling*, 151–80.

———. 2008. "Stated Preference Experimental Design Strategies." *Handbook of Transport Modelling*, 151–80.

Scholz, Michael, Verena Dorner, Markus Franz, and Oliver Hinz. 2015. "Measuring Consumers' Willingness to Pay with Utility-Based Recommendation Systems." *Decision Support Systems* 72: 60–71. https://doi.org/https://doi.org/10.1016/j.dss.2015.02.006.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28. https://doi.org/10.1257/jep.28.2.3.

This work has been accomplished with financial support of MIAI interdisciplinary institute, Grenoble, France.

Annexes

Annexe A: Econometric models

MNL and MMNL are based on the Random Utility Models (RUM) introduced and developed by McFadden (1974). Consumers optimize (and researchers estimate) an indirect utility function, that "*has a closed graph and is quasi-convex and homogeneous of degree zero in the economic variables*" (McFadden 2001). Applying the standard model to discrete choice requires the consumer's choice among the feasible alternatives to maximize conditional indirect utility based on some reference alternative, rather than absolute utility.

The functional form of the canonical indirect utility function depends on the structure of preferences, including the trade-off between different available alternatives. The perceived utility U_{ij} of individual i facing alternative $j \in \Omega$ (Ω, the set of alternatives) can be expressed as the sum of two terms: a systematic utility V_{ij} defined by some fixed deterministic function and a random residual term η_{ij} reflecting some unobserved random effects:

$$U_{ij} = V_{ij} + \eta_{ij} \tag{2}$$

Multinomial Logistic Regression (MNL)

The MNL model is one of the simplest RUM (McFadden 1974). This class of models relies on the hypothesis, that an individual i maximises his perceived utility over a set of alternatives $j \in \Omega$, as described earlier:

$$U_{ij} = V_{ij} + \eta_{ij} \text{ where } V_{ij} = \alpha_j + \beta_j X_i + \gamma Z_j \tag{3}$$

Both β, representing the alternative specific individual coefficients, and γ, standing for population-wide attributes effects, are assumed to be fixed across population, meaning that all the individuals have identical preferences and are subject to identical effects. As precised in Agresti (2013) this approach enables discrete-choice models to contain characteristics of the chooser and of the choices. More simple models may be imagined if the access to the individual characteristics or altenatives' attributes is limited, resulting in:

- $U_{ij} = \alpha_j + \beta_j X_i$ for modelling only individual characteristic effects;
- $U_{ij} = \alpha_j + \gamma Z_j$ for capturing only alternatives' attributes impacts.

The MNL is based on the assumption that the residuals η_{ij} are identically and independently distributed (iid.) as Gumbel random variables with zero mean and scale parameter σ, which is usually set to 1. The probability of choosing alternative ω_j from among those available $\{\omega_1, \ldots, \omega_k\} \in \Omega$ by individual i, can be expressed in closed form as:

$$P_{ij} = \frac{e^{V_{ij}/\sigma}}{\sum_{l=1}^{k} e^{V_{il}/\sigma}} \tag{4}$$

Mixed Multinomial Logistic Regression (MMNL)

The Mixed Logit is a further development and generalisation of MNL, because these models may be constructed using Mixed Logit specification with a correct parametrisation. The main difference from the more simple models is that in this case it is assumed that effects vary across population and might even be correlated. The utility specification in this case is constructed identically to simple models, but the deterministic part assumes that effects vary across population. Mathematically the random effects specification is achieved through the parameter vector γ_i, which is unobserved for each i. The γ in this case is assumed to vary in the population following the continuous density $f(\gamma_i \mid \theta)$, where θ are the parameters of this distribution.

$$U_{ij} = V_{ij} + \eta_{ij} \text{ where } V_{ij} = \alpha_j + \beta_j X_i + \gamma_i Z_j \tag{5}$$

The simplest choice of the distribution for the random effects is the normal distribution, which was used by Michaud, Llerena, and Joly (2012), or more precisely a multivariate normal distribution, because authors took into account the correlation between coefficients:

$$\gamma_i \sim MVN(\gamma, \Sigma) \tag{6}$$

Annexe B: CNN architecture

The designed CNN consists of two transformation layers. The first one is 1D convolutional layer with linear activation function, which takes as input the dataset in "wide" format with 27 variables overall (9 variables for each alternative), which produces a single value as an output value for each individual for each choice set, resulting in 3 output values in total. The second layer is a restricted softmax transformation layer, which directly applies softmax transformation over the inputs, without any supplementary permutations.

The vector of inputs issued from the dataset transformed into the "wide" format can be represented as:

$$X_i = Buy_{i,A}, Sex_{i,A}, Age_{i,A}, \ldots, Habit_{i,C}, Price_{i,C}, Label_{i,C}, Carbon_{i,C}, LC_{i,C} \quad (7)$$

Where $j \in \{A, B, C\}$, with C denoting the "No buy" option. All values with C index are set to zero in order to set the baseline alternative.

Figure 2: Convolution Neural Network design