Dealing with Correlations in Discrete Choice Models

THÈSE N^{O} 8170 (2018)

PRÉSENTÉE LE 23 FÉVRIER 2018
À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT
LABORATOIRE TRANSPORT ET MOBILITÉ
PROGRAMME DOCTORAL EN GÉNIE CIVIL ET ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Anna FERNÁNDEZ ANTOLÍN

acceptée sur proposition du jury:

Prof. K. Beyer, présidente du jury
Prof. M. Bierlaire, Prof. M. M. Cochon de Lapparent, directeurs de thèse
Prof. E. Cherchi, rapporteuse
Prof. C. A. Guevara, rapporteur
Prof. A. Alahi, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

"Let us pick up our books and our pens, they are our most powerful weapons. One child, one teacher, one book and one pen can change the world."

— Malala Yousafzai

To Agueda, Marc and Quim ...

Acknowledgments

It has been a bit over four years since I started this adventure. An adventure that wouldn't have been possible without so many of you. I am grateful to everyone that has helped, encouraged and made me happy during this time.

First and foremost, I would like to thank Michel Bierlaire, both from a personal and from a scientific point of view. Thank you for giving me the opportunity to join Transp-OR, first as an intern and then as a PhD student. Thank you for making me discover discrete choice and beer, and showing me that both are equally important. Thank you for guiding me throughout the thesis, for inspiring me, and for setting an example. Thank your for always finding words of encouragement, and for making everything seem simple. I admire you. From the bottom of my heart, thank you.

I am also very grateful to Matthieu de Lapparent for his co-supervision of the thesis. In particular, for his encouragement and help during the reviewing process of the first paper (Chapter 2), and for our work together in the second one (Chapter 3). This thesis has also improved significantly with the comments and feedback of the members of the jury, and for that I am grateful to them. A big thank you to Alexandre Alahi, Elisabetta Cherchi and Angelo Guevara. I am also grateful to Angelo for our work together in the first part of my PhD (Chapter 2), and for sharing his knowledge with me from the other side of the world. I also wish to thank Katrin Beyer for presiding the committee. I had imagined that the moment of defending the PhD would be a stressful one, but the six people that I mentioned above made it a pleasant and stimulating moment of discussion. Gracias, merci, grazie, danke!

Part of this thesis has been the result of a collaboration project with Nissan, and I wish to acknowledge their financial support.

I would also like to thank Pau Puig and Andreu Raig, two of my highschool teachers. I don't think I would have followed the same path in life if it weren't for them. Thanks to their passion, I discovered mine. They are the best teachers that a teenager could imagine.

I cannot imagine a better working environment than TRANSP-OR. I wish to thank all the present and past members for making the office such a relaxed and pleasant environment, for all the fun we've had in conferences around the world, and for the scientific and non-scientific discussions. Thank you Amanda, Anne, Aurélie, Antonin, Bilal, Bilge, Daniel, Dimitris, Eva, Flurin, Gael, Iliya, Jianghang, Mara, Marianne, Mila,

Nicholas, Nikola, Nitish, Shadi, Sofia, Stefano B., Stefano M., Xinjun, Yousef, Yuki, Zhengchao. Also the students that I have had the pleasure to work with in different projects. You have all taught me something: Anna-Katharina, Gabriel, Martí, Maurin, Michael, Nicola, Takao and Thibaut. Some colleagues have become very important people in my life. Thank you, Tom, for making our apartment feel home; Riccardo, for your friendship; Meri, for all the discussions where it is us against the world, for making me discover Youtube, for setting up the framework that I have used in Chapter 4, and for sharing with me your code; Virginie, for all your professional and non-professional help. For our work together in Chapter 4, for reading the different versions of my thesis three thousand times, for guiding me during the stressful time at the end of the PhD, but, more importantly, for your bravery and for showing me that it is worth it to follow your dreams. You'll be an excellent professor soon, I'm sure! The Transp-OR family also goes further than Transp-OR. Thank you Aitor, Sophia, Darko and Christina for being part of this adventure.

The most important thing that I have learned during the PhD is that distance is not necessarily measured in kilometers. This work wouldn't have been possible without everyone back home, that have felt so close since I left. Gracias, Pat, por estar siempre, incondicionalmente, para todo. Gràcies, Roger, per fer-me saber que sempre puc comptar amb tu. Gràcies, Helena, per totes les trobades als diferents llocs en els que hem viscut durant aquest anys. Gràcies Ona i Sílvia, per ser part de la famlia, i gràcies Jordi i Enric per fer-les felices. Gràcies Rym i Bebè, que sense haver nascut ja em feu feliç! I want to thank Queralt, Ingrid, Albert, Sergi, Oscar B., Oscar C., Raquel and Aida, for having come to visit me to Lausanne. I am also thankful to those that have been ready to change their schedule when I was calling 10 minutes in advance saying I was home: Anneta, Juancho, Andrea, Guille, Sílvia and Heather.

Today, I am happy to say that I have two homes: Barcelona and Lausanne. I am grateful to everyone that has made me feel home here. I want to thank all my waterpolo coaches: Jaap, Seb, Marc, Alexis, Adam and Claudio, and all the members of my waterpolo teams: Lausanne 2, Lausanne 3 and Nyon. I am so lucky for all the people I have met in or around the swimming pool. Among them, I would like to thank particularly Lara, Kristelle, Christelle and Lise. I would also like to thank my Catalan crew in Lausanne, as well as all the rest of my friends from Lausanne, the rest of Switzerland and around the world. I am also grateful to my new family in Switzerland and Austria for being so welcoming.

I would also like to thank my family, as well as Pepi and Pedro. Però sobretot a vosaltres, Mama, Papa i Marc. Tot el que tinc que agrair-vos no hi cabria aquí. Gràcies per acceptar la meva decisió de marxar fora, i per fer veure que no ens trobem a faltar, pequè Lausanne i Barcelona no estan tan lluny. A tu, Marc, per ser el millor germà i amic que es pot tenir. A tu, Papa, per ensenyar-me que no s'ha de perdre mai l'humor. La vida és com és, però no per això deixarem de riure ni de passar-nos-ho bé. I a tu, Mama, per ser la dona forta que ets, i per ensenyar-me a ser una dona forta a mi. Crec

que és impossible tenir una infància i una adolescència més feliç de la que m'heu donat.

Et finalement, merci, Stefan, pour partager ta vie avec moi, pour me rendre plus heureuse, et pour m'aimer.

Lausanne, 8th December 2017 Anna Fernández Antolín

Abstract

The focus of this thesis is to develop methods to address research challenges related to correlation patterns in discrete choice models. In the context of correlations within alternatives, we extend the novel methodology of the multiple indicator solution (MIS) to deal with endogeneity, and show, through its theoretical derivation, that it is applicable when there are interactions between observed and unobserved variables. In the context of correlations between alternatives, we discuss the importance of using models that can capture them, such as cross nested logit models. We show, through real world examples, that ignoring these correlation patterns can have severe impacts on the obtained demand indicators, and that this can lead to wrong decisions by practitioners. We also address the challenge of using revealed preference data, where the attributes of the non-chosen alternatives are unavailable, and propose a solution based on multiple imputations of their empirical distributions.

In the thesis, we also contribute to the existing literature by gaining a better understanding of private motorized modes, in terms of modal split and purchases of new cars. Related to modal split, we use a mode choice case study in low density areas of Switzerland. We find that ignoring the *car-loving* attitude of individuals leads to incorrect value of time estimates and elasticities, which might have severe implications in the pricing schemes of public transportation, for example. Related to the purchase of new cars, we use data from new car acquisitions in France in 2014, and focus on hybrid and electric vehicles. We find elasticities to price that are in line with the literature, and willingness to pay values in line with the market conditions. We also study the impact of different future policy scenarios and find that the sales of new electric vehicles could reach around 1% as a result of a major technological innovation that would render electric vehicles less expensive.

In the last part of the thesis, we propose the *discrete-continuous maximum likelihood* (DCML) framework, which consists in estimating discrete and continuous parameters simultaneously. This innovative idea, opens the door to new research avenues, where decisions that were usually taken by the analyst can now be data driven. As an illustration, we show that correlations between alternatives can be identified at the estimation level, and do not need to be assumed by the analyst. The DCML framework consists in a mixed integer linear program (MILP) in which the log-likelihood estimator is linearized. This linearization might be useful to estimate parameters of other discrete

choice models for which the log-likelihood function is not concave (and therefore global optimality is not insured by the optimization algorithms), since for an MILP, a global optimum is guaranteed. We use a simple mode choice case study for the proof-of-concept of the DCML framework, and use it to investigate its strengths and limitations. The preliminary results presented in the thesis seem very promising.

To summarize, we develop methods to deal with correlations in discrete choice models that are relevant to real world problems, and show their applicability by using transportation examples. The contributions are therefore both theoretical and applied. The new methods proposed open the door to new research directions in the discrete choice field.

Keywords: Mathematical modeling of behavior, discrete choice models, endogeneity, multiple indicator solution, latent variables, revealed preference data, nested logit, cross nested logit, discrete-continuous maximum likelihood, mixed integer linear program, log-likelihood linearization, multiple imputations, transportation mode choice, car-type choice, value of time, policy analysis, willingness-to-pay, elasticities, electric vehicles, hybrid vehicles.

Résumé

Le but de cette thèse est de développer des méthodes permettant de tenir compte des différentes corrélations pouvant exister au sein de modèles de choix discrets. Lorsque la corrélation est interne aux alternatives, on parle d'endogénéité. Dans la première partie de cette thèse, nous étendons une nouvelle méthodologie, appelée solution des indicateurs multiples, qui sert à corriger cette endogénéité. Nous montrons, au travers de sa dérivation théorique, que nous pouvons appliquer la méthodologie des indicateurs multiples lorsqu'il existe des interactions entre des variables observées et non observées. Dans la deuxième partie de cette thèse, nous montrons qu'il est également important de tenir compte des corrélations qui existent entre les différentes alternatives et nous montrons, à l'aide d'exemples, qu'ignorer ces corrélations peut avoir de sérieux effets sur les indicateurs de demande obtenus, avec comme conséquence ultime une mauvaise prise de décision. Les données utilisées dans cette thèse sont des préférences révélées, où les attributs des alternatives non choisies ne sont pas disponibles. Afin d'obtenir des attributs pour ces alternatives, nous proposons comme solution d'avoir recours à l'imputation multiple sur base de leur distribution empirique.

Dans cette thèse, nous contribuons aussi à la littérature existante grâce à une meilleure compréhension des modes de transport motorisés privés, en particulier du côté de la distribution modale et des achats de nouveaux véhicules. En ce qui concerne la distribution modale, nous utilisons une étude de cas située dans les zones à faible densité de population en Suisse. Nos résultats montrent qu'ignorer l'attitude de *préférence pour la voiture* produit des valeurs du temps et des élasticités incorrectes, ce qui peut avoir de sérieuses implications sur les régimes de prix des transports publics, par exemple. Concernant l'achat de nouvelles voitures, nous utilisons des données d'achats de nouvelles voitures en France en 2014, et nous nous concentrons sur les voitures électriques et hybrides. Nous obtenons des élasticités de prix qui sont alignées avec la littérature existante, et des valeurs du consentement à payer alignées avec les conditions du marché. Nous étudions aussi l'impact de différents scénarios futurs, et constatons que les ventes de nouvelles voitures électriques pourraient s'élever à environ 1%, à la suite d'une innovation technologique qui les rendrait moins chères.

Dans la dernière partie de cette thèse, nous proposons un cadre pour le *maximum de vraisemblance discret et continu*, qui consiste à estimer des paramètres discrets et continus simultanément. Cette idée novatrice ouvre la porte à de nouvelles directions de

recherche, où certaines décisions qui sont d'habitude prises à priori par l'analyste, peuvent maintenant être révélées par les données. Comme exemple, nous montrons comment les corrélations entre alternatives peuvent être identifiées à l'étape de l'estimation, et ne doivent ainsi pas être supposées par l'analyste. Ce cadre consiste en un problème linéaire à variables mixtes (PLM) dans lequel l'estimateur de log-vraisemblance est linéarisé. Cette linéarisation peut être utile pour estimer des paramètres d'autres modèles de choix discret pour lesquels la fonction de log-vraisemblance n'est pas concave, puisque l'optimalité globale est garantie dans un PLM. Nous utilisons un cas d'étude de choix modal simple pour prouver que le cadre fonctionne, et pour étudier ses avantages et ses limitations. Les résultats préliminaires présentés dans la thèse sont très encourageants.

En conclusion, cette thèse propose des méthodes pour adresser le défi de la présence de corrélation dans des modèles de choix discret. Nous montrons l'applicabilité de ces méthodes sur des exemples issus du milieu des transports. Les contributions de la thèse sont donc théoriques et appliquées. Les nouvelles méthodes proposées ouvrent la porte à de nouvelles directions de recherche dans le domaine du choix discret.

Mots clés : Modélisation mathématique du comportement, modèles de choix discret, endogénéité, solution des indicateurs multiples, variables latentes, données de préférence révélées, logit imbriqué, logit imbriqué croisé, maximum de vraisemblance discret et continu, problème linéaire à variables mixtes, linéarisation de la log-vraisemblance, imputations multiples, choix du mode de transport, choix du type de voiture, valeur du temps, analyse de politiques, consentement à payer, elasticité, voitures électriques, voitures hybrides.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context and motivation

Human behavior is studied in many different fields such as psychology, psychiatry, sociology and anthropology. It can also be analyzed from a mathematical and economical point of view, using mathematical models to understand and to predict how people make choices. Demand modeling is the tool used to study the decision making process of people. The result of a decision-making process can either be continuous or discrete. To model continuous variables (e.g.: quantity of water consumed per year in a household), regression models are usually applied. When the outcome of a decision is discrete (e.g.: if a car is going to be purchased in the following month or not), discrete choice models are utilized. Two of the main advantages of discrete choice models (DCM) are that they are *probabilistic* and *disaggregate*. They are probabilistic, meaning that the output that they provide are the choice probabilities of each alternative. They are disaggregate, meaning that these probabilities are individual specific, expressed as a function of their socioeconomic characteristics. In other words, DCM provide the probabilities for each individual to choose each alternative.

DCM have experienced an increase of their popularity after professor McFadden received the Nobel Prize in the year 2000 for his development of theory and methods for analyzing discrete choice. The Nobel prize website states that professor McFadden "Showed how to statistically handle fundamental aspects of micro-data, namely data on the most

important decisions we make in life."[1] Indeed, DCM can be used to address challenges related both to the public and the private sectors. With regard to the private sector, companies might use discrete choice models to determine which is the optimal price of an alternative to maximize the revenue, or to determine which market segment is more responsive to a change in price. With regard to the public sector, DCM can help governments and other organizations to find answers to questions related to the well being of society, such as those related to climate change (e.g.: which policies are more effective to increase the market shares of hybrid and electric vehicles).

Recent applications of DCM in the public sector include the study of social welfare and tax distribution by governments (Ozdemir et al., 2016), health economics (N. Flynn et al., 2013), evacuation decisions (Sadri et al., 2017; Hasan et al., 2011) and environmental economics (Webb et al., 2017; Rodrigues et al., 2016), to name a few. Recent applications in marketing include sponsorship decision making (Johnston and Paulsen, 2014), pricing plans for financial advisory services (Schlereth, 2014), understanding television channel search and commercial breaks (Yao et al., 2016), and mobile advertisement (Bart et al., 2014), among others. Chorus (2015) opens the door to applying DCM to model moral decisions.

If DCM are so broadly applied, it is thanks to the progress that has been done in this field in last decades. The logit model is the element from which the rest of the models build on. Its key advantage relies on the simplicity of its closed-form probability expression. In the logit model, the error terms are supposed to be independently and identically distributed (iid) across alternatives and individuals. This leads to the independence of irrelevant alternatives (IIA) property, which states that the ratio of two choice probabilities is independent of the attributes or even the existence of any other alternatives. This may imply unrealistic substitution patterns across alternatives. To relax this assumption, more advanced models have been proposed. Among these, are the nested and the cross nested logit models. The common idea behind them is to allow correlations between alternatives, by placing them into several groups called nests. Alternatives that belong to the same nest share a common error term and are therefore correlated, while alternatives that are not in the same nest are independent. Another extension to the previously discussed models, includes the mixed logit models, where some of the estimated parameters are supposed to be a random variable that follows a certain distribution. Mixtures of models can be used, for example, to model the stochasticity in the sensitivity of travel time among the population. Mixtures can also be used to mimic nesting structures, by adding common error terms to alternatives that are assumed to be in the same nest. A particular case of mixtures of models are hybrid choice models, by which attitudes and perceptions can be included in the utilities of the alternatives. Attitudes (e.g.: *environmental friendliness*), might play an important role in some choices (e.g.: the purchase of a new car), and can not be modeled by using only

[1] http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2000/mcfadden-facts.html

observable characteristics of an individual. To take these attitudes and perceptions into account, psychometric indicators are also collected.

By using the models described above, demand indicators can be derived. First, they can be used to forecast the market shares of the alternatives in several future scenarios. Second, elasticities can be calculated. They are used to analyze the variation of probabilities as a consequence of a variation of an attribute. For instance, it might be interesting to study how is the market share of private motorized modes affected by an increase of fuel prices: if the fuel cost increases by 1%, by which percentage is the total number of car trips affected? Third, the willingness to pay, which translates the sensitivity towards an attribute to its monetary value, can also be derived. By computing willingness to pay indicators, the following questions can be answered: how much are people willing to pay in order to save one minute of travel time from their origin to their destination? Does this value depend on the mode? Does it vary with the income level of the person?

However, DCM are very data demanding. In order to apply them, we need, for each individual, (i) the considered choice set, (ii) the attributes of the chosen alternative, (iii) the attributes of the non-chosen alternatives, and (iv) a list of socioeconomic characteristics. Two types of data exist, stated preferences (SP) and revealed preferences (RP). In SP data, the modeler first designs and creates a survey and then faces the respondents to one or several hypothetical choice tasks. This data is easier to obtain, but has more limitations, that are discussed later on in the thesis. RP data, on the other hand, is more challenging to obtain, but more realistic. It describes actual choices that people have done. Often, the information on the choice set and the attributes of the unchosen alternatives are missing. This is true, in particular, in new types of data such as those collected from mobile applications. It is crucial to the field that we learn how to use this data. Steps in this direction are discussed in Chapter 3. We use an RP dataset where the unchosen alternatives and its attributes are not provided. We use aggregation techniques for the choice set definition and we therefore have to find a way to impute their attributes. It is important that the applicability of the methods developed is tested using RP data. The values in the SP datasets are engineered by the modeler, so even if the methods work from a theoretical point of view, they might fail in practice, when the data used corresponds to observed choices.

Moreover, there are also methodological challenges. Even when the data is accurate, it might be incomplete. Omitting important attributes from the model specification leads to correlation between the deterministic and the stochastic parts of the model. This is called *endogeneity*. Challenges related to correlations within alternatives are addressed in Chapter 2, by extending an existing methodology to address endogeneity, and applying it to revealed preference data.

There might also be correlations between alternatives. Two alternatives are correlated if they share unobserved attributes. Standard modeling techniques often ignore these

correlations, generally because it is not clear which alternatives are correlated. An example of this can be found in airline itinerary choice models, where alternatives can be correlated due to the time of departure, the number of connections, the operating airline, or the airport where a connection will take place, to name a few dimensions.

The omission of a latent (or unobserved) variable might cause the two types of correlations mentioned above. If the *environmental friendliness* of a person is not taken into account in a car-type choice case study, we might observe endogeneity, since the environmental friendliness of a person might affect her willingness to pay towards alternative fuel vehicles. At the same time, not including this variable in our models might also cause that the alternative fuel vehicle alternatives are correlated. In both cases, ignoring the correlation patterns (within and between alternatives) generates wrong forecasts. This might result in companies or governments taking wrong decisions. Methodological advances in this direction are therefore necessary. These challenges are addressed in Chapters 3 and 4. In Chapter 3, we assume a nesting structure to take into account the correlation between alternatives. This motivates the next chapter, where we aim at not having to assume a nesting structure, but to estimate it together with the parameters of the model, instead.

In conclusion, modeling correlations within and across alternatives is imperative to obtain accurate demand indicators. The specific objectives and contributions for each part are summarized in the following section.

In the context of this thesis, we focus on DCM applied to transportation, with a special interest in private motorized modes (PMM). According to the European Commission, transport is the only major sector in the EU for which greenhouse gas emissions are not decreasing[2]. Personal transportation, with the emissions from cars and vans, represents around 15% of the total EU emissions of CO2. Bearing this in mind, it is important for governments and policy makers to decrease the modal split of private motorized modes, and to shift from internal combustion engines towards hybrid and electric vehicles. In order to understand better the modal split between PMM, public transportation (PT) and slow modes (such as walking and biking) we study how the omission of the *car loving* dimension leads to inaccurate demand indicators. We also focus on the market shares of hybrid and electric vehicles, and how different future policy scenarios affect their sales. Finally, we use an SP dataset collected to decide whether or not to build a new transportation system in Switzerland. The three case studies used in this thesis are briefly described below.

PostBus: the purpose of this case study is to analyze travel behavior in low density areas of Switzerland. The dataset was collected between 2009 and 2010 in a joint collaboration between EPFL and PostBus (a major Swiss bus operator). It consists of

[2]https://ec.europa.eu/clima/policies/transport/vehicles_en

revealed preference mode choice data, as well as psychometric indicators. It is used as a case study in Chapter 2.

New car sales: this case study aims at modeling the purchases of new cars. It is based on a revealed preference survey of new car purchases, which is representative of the 2014 French car market. The data was provided by Nissan SA in the framework of a joint project. It is used in Chapter 3.

Swissmetro: the aim of this case study is to analyze the impact of a modal innovation in transportation: the Swissmetro, which was thought as a revolutionary mag-lev underground system. The dataset consists on survey data collected in 1998. Since the mode did not exist, stated preference data had to be collected. We use this dataset to illustrate the methodology developed in Chapter 4.

The three datasets described above were collected in order to find answers to real-world problems. This illustrates how useful discrete choice models are in practice.

1.2 Research objectives

This thesis aims at proposing solutions to some of the research challenges of the state-of-the-art discrete choice models. In particular, the objectives are:

1. **Correlation within alternatives (endogeneity):** to show that the multiple indicator solution method is applicable when there are interactions between observed and unobserved factors in the specification of the utility function.
2. **Revealed preference data:** to define an operational way to impute the attributes of the non-chosen alternatives.
3. **Correlation between alternatives:** to find a way to automatically determine which alternatives share unobserved correlations, by estimating the discrete parameters that allocate alternatives to nests, and the continuous parameters of the model simultaneously.
4. **Application:**
   (a) to apply the developed methodologies to revealed or stated preference data so as to evaluate their performance and applicability,
   (b) to shed light about if and how the *car loving* dimension affects the modal split and its related demand indicators, such as elasticities and value of time.

(c) to gain insights in the new car market in France in 2014. We focus mainly on hybrid and electric vehicles.

1.3 Scientific contributions

We seek to contribute towards developing methods to address both data, and methodological research challenges due to correlation patterns. In particular:

1. **Correlation within alternatives (endogeneity):**
   - We develop the theoretical derivation to show that the multiple indicator solution (MIS) method can also be used to account for endogeneity when there are interactions between observed and unobserved factors.
   - We apply the MIS method for the first time to RP data using a mode choice case study.
2. **Revealed preference data:**
   - We use RP data from French new car market in 2014 to analyze and compare different policy scenarios, price elasticities and willingness to pay and to accept.
   - We propose to impute the attributes of the non-chosen alternatives by drawing from their empirical distributions using multiple imputations.
   - We use a cross nested logit model to take into account the unobserved attributes that different car-types share.
3. **Correlation between alternatives**
   - We introduce the concept of *discrete-continuous maximum likelihood*, in order to automate the process of estimating discrete parameters. We apply it to identify which alternatives share unobserved attributes.
   - We apply the developed framework to a simple case study to show its applicability and discuss its potentials and limitations.
   - We linearize the log-likelihood function by relying on simulation.
   - We generalize the framework and show that in can be used in other applications thanks to the linearization of the log-likelihood function.

1.4 Structure

The thesis is structured as follows.

Chapter 2 describes the proposed extension of the MIS methodology to deal with endogeneity and applies it to a mode choice case study with revealed preference data.

The literature review in the chapter borrows from:

> Fernández-Antolín, A., Stathopoulos, A., and Bierlaire, M. (2014). Exploratory Analysis of Endogeneity in Discrete Choice Models. Proceedings of the 14th Swiss Transport Research Conference (STRC) 14-16 May, 2014.

Preliminary ideas of the methodology presented are published as:

> Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M. (2015). Correcting for endogeneity using the EMIS method: a case study with revealed preference data. Proceedings of the 15th Swiss Transport Research Conference (STRC) 15-17 April, 2015.

> Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity using the multiple indicator solution. Technical report TRANSP-OR 160405. Transport and Mobility Laboratory, ENAC, EPFL.

The work has been presented in preliminary stages in the following conferences:

- 4th Symposium of the European Association for Research in Transportation (hEART), Technical University of Denmark, September 09, 2015, Copenhagen, Denmark
- 14th International Conference on Travel Behaviour Research (IATBR), July 19, 2015, Beaumont Estate, Windsor
- Workshop on Discrete Choice Models 2015, May 28, 2015, Lausanne, Switzerland
- 15th Swiss Transport Research Conference (STRC), April 16, 2015, Monte Verità, Ascona, Switzerland
- 3rd Symposium of the European Association for Research in Transportation (hEART), Institute for Transport Studies, University of Leeds, September 11, 2014, Leeds, United Kingdom
- 14th Swiss Transportation Research Conference, May 15, 2014, Monte Verità, Ascona, Switzerland

The chapter has been published as:

> Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data, *Journal of Choice Modelling 20:1-15*.

Chapter 3 proposes a cross nested logit structure to model purchases of new cars. Multiple imputations are used for the attributes of the unchosen alternatives. Elasticities, willingness to pay and market shares in future possible scenarios are investigated.

The preliminary work related to this chapter is published as:

> Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M. (2016). Uncovering substitution patterns in new car sales using a cross nested logit model. Proceedings of the 16th Swiss Transport Research Conference (STRC) 18-20 May, 2016.

And the chapter is based on the article:

> Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M. (2017). Modeling purchases of new cars: an analysis of the 2014 French market. Accepted for publication in the journal *Theory and decision*.

Preliminary stages of this work have been presented in the following conferences:

- 5th Symposium of the European Association for Research in Transportation (hEART), Delft University of Technology, September 14, 2016, Delft, Netherlands
- TRISTAN IX, June 13, 2016, Oranjestad, Aruba
- 16th Swiss Transport Research Conference (STRC), May 19, 2016, Monte Verità, Ascona, Switzerland
- Workshop on Discrete Choice Models 2016, April 22, 2016, Lausanne, Switzerland

Chapter 4 introduces the concept of discrete-continuous maximum likelihood, used to determine simultaneously the parameters of a nested logit model and the best nesting structure. The methodology is illustrated with a mode choice case study.

Preliminary ideas related to the chapter are published as

> Fernández-Antolín, A., Lurkin, V., de Lapparent, M., and Bierlaire, M. (2017). Discrete-continuous maximum likelihood for the estimation of nested logit models. Proceedings of the 16th Swiss Transport Research Conference (STRC) 17-19 May, 2017.

The chapter is based on the article

> Fernández-Antolín, A., Lurkin, V., and Bierlaire, M. (2017). Discrete-continuous maximum likelihood for the estimation of nested logit models. Working paper.

Previous stages of the work in this chapter have been presented in the following conferences:

- 2017 INFORMS Annual Meeting, October 23, 2017, Houston, USA
- 6th Symposium of the European Association for Research in Transportation (hEART), Technion, September 12, 2017, Haifa, Israel
- Workshop on Discrete Choice Models 2017, EPFL, June 23, 2017, Lausanne, Switzerland
- 16th Swiss Transport Research Conference (STRC), May 18, 2017, Monte Verità, Ascona, Switzerland

Chapter 5 summarizes the contributions of the thesis and determines future research directions.

2

Correcting for endogeneity due to omitted attitudes

This chapter is based on the article:

Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data, *Journal of Choice Modelling* 20:1-15. doi:10.1016/j.jocm.2016.09.001

The work has been performed by the candidate under the supervision of prof. C. Angelo Guevara, prof. Matthieu de Lapparent and prof. Michel Bierlaire.

2.1 Introduction

Endogeneity is an issue that often arises in demand modeling. One of the assumptions to derive random utility models such as logit, probit, nested logit and cross nested logit is that the deterministic part of the utility function is independent from unobserved factors. If this assumption is violated, it may result in inconsistent estimates of the parameters. This is what is known as endogeneity. As Guevara (2010) describes, it can have three main causes: (i) errors in the measurements of the variables, (ii) simultaneous

determination and (iii) omitted variables.

The first cause is very intuitive: if there are systematic errors in the measurements, these propagate to the error term, which is then correlated with the wrongly measured variable. An example of the second cause in the context of transportation can be found in the simultaneous modeling of mode and housing choice. People with a tendency to travel by public transportation locate closer to stations, thus making their travel times shorter on that mode. The residential location choice is affected by the mode choice, but at the same time the mode choice is affected by the residential location choice. This is known as simultaneous determination, and assuming that one is an exogenous explanatory variable would result in a wrong measurement of their true impact on one another.

An example of the third cause can also be found in transportation, when an unobserved variable - such as comfort - is not included in the model. In a mode choice between public transportation and private modes, assume that there is an observed attribute (travel time, travel cost) that is correlated with an unobserved attribute (perception of comfort). If comfort is omitted, we may obtain biased estimates for the parameters associated with time and/or cost. This can be seen intuitively as follows: if people are traveling at peak hours when public transportation is very congested, the disutility towards public transportation caused by discomfort is captured by the travel time parameter. It results in a downwards-estimated parameter for travel time, since it captures both the disutility towards public transportation caused by travel time and the disutility caused by discomfort. In a similar way, transportation systems that are more expensive because they are more comfortable - like traveling in the first class in a train - have an upwards estimated parameter related to cost. This parameter is capturing on the one hand the disutility for high prices, but on the other hand the fact that travelers are willing to pay higher prices to travel in a more comfortable way. It can even result in positive estimates for parameters related to cost. This results, of course, in wrong willingness to pay estimates.

The problem may appear as well when a latent variable is omitted. There is evidence in the literature that *car lovers* have a different value of time for private motorized modes compared to other individuals who don't have this preference (Atasoy et al., 2013). If this is not explicitly modeled, the estimator of value of time may not be consistent. In terms of the specification of the utility, there is evidence in the literature that suggests to use the interaction of *car lovingness* and cost to address heterogeneity of taste (Abou-Zeid et al., 2010).

As discussed above, endogeneity can yield to biased and inconsistent estimates. However, it is rarely assessed and corrected for in practical applications. This is due to the fact that although several methods to correct for it exist (BLP, control function...), they rely on instruments, that are not straightforward to identify in practice. A complete review

of these methods is found in Section 2.2. In this chapter, we build on the Multiple Indicator Solution (MIS), that can be applied when there is an interaction between the unobserved factor and a measurable variable. We show that it can be generalized to models with interactions between observed and unobserved factors. Moreover, it is the first application with revealed preference (RP) data of the MIS method, that has only been tested with stated preference (SP) data (Guevara and Polanco, 2016). We apply the MIS methodology in order to get more realistic value of time (VOT) and time elasticity estimates from a mode choice revealed preference dataset in Switzerland. The values of time obtained with the corrected model are able to account for its heterogeneity in terms of the latent attitudes toward the car, or the degree of car-lovingness. We show that the MIS handles correctly the endogeneity issue by comparing it with the integrated choice and latent variable (ICLV) approach, which is assumed to give a full account of the data generation process, but at a significantly higher computational cost.

Qin (2015) highlights that, "heated methodological debates over the causal validity and credibility of instrumental variable-based estimates" have arisen over the last decades. We also refer to Angrist et al. (1996) and Imbens (2014), for extensive summaries of the IV approach. As the MIS estimation method is taking inspiration from the instrumental variable (IV) estimation method, similar problems apply to the MIS estimation method. The choice of appropriate indicators that satisfy relevance and (conditional) exogeneity assumptions is very important. It however depends on the application, as is discussed later in the chapter.

This chapter is structured as follows: the literature review is presented in Section 2.2, followed by the description of the theoretical framework in Section 2.3. Section 2.4 contains the case study, along with a discussion of the results obtained. Finally, the conclusions and future work directions are discussed in Section 2.5.

2.2 Literature Review

This section is divided in two subsections: Section 2.2.1 is a detailed review of the different methodologies that have been proposed in the literature to address endogeneity. Section 2.2.2 gives some insight in the existing literature related to modeling attitudes and perceptions.

2.2.1 Endogeneity

Louviere et al. (2005) present the recent progress that has been done in the field of endogeneity in discrete choice models. However, they give a very broad definition of endogeneity and focus also on choice set formation, interactions among decision makers

and models of multiple discrete/continuous choice amongst other topics. In this review, as well as during the whole thesis, we are going to focus only on how to correct for endogenous explanatory variables, in the same sense considered by Guevara (2015).

A widely used methodology is the BLP (Berry et al., 1995, 2004) named after its authors. This approach consists in removing the endogeneity from the non-linear choice model and dealing with it in linear regressions. This requires adding an alternative specific constant (ASC) for each product and each market, the level in which the endogeneity problem is assumed to occur. A description of the instrumental variable methodology can be found in most of the basic econometric textbooks such as Baum (2006), Lancaster (2004) or Wooldridge (2010). Guevara (2010) describes in his thesis why it is more complex to deal with endogeneity in discrete choice models compared to linear models: these corrections lead to changes in the error term which imply a change of scale in the discrete choice models.

There are many studies that use the BLP approach to deal with endogeneity in discrete choice models. To name some examples, Walker et al. (2011) introduce a social influence variable in a behavioral model which is endogenous, as the factors that impact the peer group also influence the decision maker and this causes correlation between the field effect variable and the error. Train and Winston (2007a) use the BLP approach to correct for price endogeneity in automobile ownership choice. Crawford (2000) uses it for consumers' choice among TV options and Nevo (2001) uses it for a study of the cereal industry. It is also the approach chosen by Goolsbee and Petrin (2004) where they examine the direct broadcast satellites as a competitor to cable TV.

A second approach in the literature is the control function methodology. The concept dates back to Hausman (1978) and Heckman (1978), although the term *control function* was introduced by Heckman and Robb Jr. (1985). Petrin and Train (2009) describe a control function approach to handle endogeneity in choice models. They apply both the control function and the BLP methodologies in a case study and find similar and more realistic demand elasticities than without correcting for endogeneity. They describe the control function methodology in detail. Guevara (2010) also uses this method to study the choice of residential location. He also shows that there is a link between the control-function methods and a latent-variable approach. Both the BLP and the control function method rely on two steps estimation procedures with instrumental variables.

The third frequently used approach is the one that Guevara (2010) calls *the control-function method in a maximum-likelihood framework* and Train (2003) calls *maximum-likelihood method*. It is the same formulation used by Villas-Boas and Winer (1999) in brand choice models and Park and Gupta (2012). In particular, Park and Gupta (2012) propose what they describe as a "new statistical instrument-free method to tackle the endogeneity problem". They model the joint distribution of the endogenous regressor and the structural error term by a Gaussian copula and use nonparametric density

estimation to construct the marginal distribution of the endogenous regressor. Also, Bayesian methods to handle endogeneity have been introduced by Yang et al. (2003) and Jiang et al. (2009).

Endogeneity can also be mitigated by the Integrated Choice and Latent Variable (ICLV) approach, where a latent factor captures an unobserved qualitative attribute. This methodology explicitly models attitudes and perceptions using psychometric data. For the estimation of the parameters, maximum likelihood techniques are used, which lead to complex multi-dimensional integrals. Thus, it is a computationally intensive method.

A more novel method used for discrete choice models is the Multiple Indicator Solution (MIS) which is described by Wooldridge (2010) in the context of linear models and generalized by Guevara and Polanco (2016) for discrete choice. As opposed to the control-function method, the MIS method does not need instrumental variables. Instead, it uses indicators to introduce a factor of correction in the choice model in order to obtain consistent estimators. Its performance is compared using Montecarlo experiments to other methodologies in Guevara (2015).

Many other methods to correct for endogeneity exist. For example, the analogous to the standard 2-stage instrumental variable approach used in regression, described by Newey (1985) does not provide correct estimates of the aggregate elasticities of the models. Guevara (2010) shows it with a case study. Another method, developed by Amemiya (1978), is as efficient as the control function approach, as shown by Newey (1987), and is globally efficient under some circumstances, but is much more complex to calculate because it involves the estimation of auxiliary models.

2.2.2 Attitudes and perceptions

A lot of literature also exists in how attitudes, perceptions and psychological factors in general play an important role in the modeling of behavior. A non-exhaustive list of research related to this would include Ajzen (2001); Olson and Zanna (1993); Wood (2000); McFadden (1986); Ben-Akiva and Boccara (1987). In particular, there are several studies describing the role of attitudes and perceptions in mode choice, such as Koppelman and Hauser (1978); Proussaloglou and Koppelman (1989); Golob (2001); Outwater et al. (2003); Vredin Johansson et al. (2006). Walker (2001) develops the most commonly used framework to include these in discrete choice models: the integrated choice and latent variable approach. However there had already been some developments of latent variable models prior to her work, such as Everitt (1984); Bollen (1989).

An interesting measure that can be derived from mode choice models is the value of time (VOT), that is defined as the amount of money that users are willing to pay to save one unit of travel time. In other words, it is the trade-off that users consider between

the time that they spend traveling and the amount of money that they are willing to pay. The first person to introduce the concept of value of time in travel behavior was Dupuit (1844, 1849). The VOT varies across individuals and the trips, characterized by variables such as age, gender, income, trip purpose... It can also be distributed (see, among others, Ben-Akiva et al. (1993); Fosgerau (2006); Hess and Axhausen (2004)).

An attitude that has been considered relevant for the estimation of the VOT is the *car loving* attitude (Abou-Zeid et al., 2010; Atasoy et al., 2013). *Car lovers* are defined as people that have an intrinsic preference towards car, for many reasons, including convenience, reliability, and symbol of social status. If either the time or the cost are actually interacting with the attitude, and it is omitted in the model specification, it then enters the error term, causing endogeneity.

2.3 Methodology

This section introduces the methodology that is used in the chapter. Section 2.3.1 is an introduction to the Multiple Indicator Solution (MIS) method. The following sections investigate how to adapt this methodology to capture possible interactions between observed attributes and unobserved factors. Section 2.3.2 contains the derivation of an intuitive but not useful approach, while Section 2.3.3 proposes a way to overcome the limitations of the previous approach. Finally, Section 2.3.4 is a reminder of the Integrated Choice and Latent Variable (ICLV) framework, that is used as a benchmark for the MIS with interactions in the case study.

2.3.1 MIS method

The multiple indicator solution method was introduced by Wooldridge (2010) for linear models and extended to discrete choice models by Guevara and Polanco (2016). It can be summarized as follows.

Consider a setup where the choice of an alternative i by a decision-maker n depends on an economic factor t_{in}, an unobserved attribute q_{in} that is correlated to t_{in}, and on a set of other explanatory variables x_{in}. The utility function of this alternative is specified as follows

$$U_{in} = \text{ASC}_i + \beta_x x_{in} + \beta_t t_{in} + \beta_q q_{in} + e_{in}, \tag{2.1}$$

where ASC_i, β_x β_t and β_q are parameters to estimate and e_{in} is a random error term. If the term $\beta_q q_{in}$ is omitted, it would enter the error term. Therefore, the error term would be correlated to t_{in} causing endogeneity. We assume that we have two indicators

I_{1in} and I_{2in} which are related to the omitted variable q_{in}. The following relation can be defined

$$I_{1in} = \alpha_0 + \alpha_q q_{in} + e_{I_{1in}}, \tag{2.2}$$
$$I_{2in} = \delta_0 + \delta_q q_{in} + e_{I_{2in}}, \tag{2.3}$$

where the pairs of variables (q, e_{I_1}),(x, e_{I_1}), (q, e_{I_2}), (x, e_{I_2}),(e_{I_1}, e_{I_2}) are independent[3], $\alpha_q \neq 0$ and $\delta_q \neq 0$. x represents the vector of explanatory variables in Equation (2.1). From Equation (2.2) we obtain $q_{in} = (I_{1in} - \alpha_0 - e_{I_{1in}})/\alpha_q$. By substituting this expression in Equation (2.1) and denoting $\theta_q = \frac{\beta_q}{\alpha_q}$ we obtain

$$U_{in} = ASC_i + \beta_t t_{in} + \beta_x x_{in} + \theta_q I_{1in} - \theta_q \alpha_0 - \theta_q e_{I_{1in}} + e_{in}. \tag{2.4}$$

The above model is still endogeneous since I_{1in} is correlated with $e_{I_{1in}}$. We therefore apply the control function method (similarly as in Guevara (2010)) and use I_{2in} as an instrument for I_{1in}. This can be done because both indicators are correlated, and I_{2in} is independent of $e_{I_{1in}}$. We can therefore define the following relations

$$I_{1in} = \gamma_0 + \gamma_1 I_{2in} + \gamma_t t_{in} + \gamma_x x_{in} + \delta_{in}, \tag{2.5}$$
$$e_{I_{1in}} = \beta_\delta \delta_{in} + \nu_{in}, \tag{2.6}$$

where δ_{in} captures the part of $e_{I_{1in}}$ which is correlated with I_{1in} and ν_{in} is an exogenous error term.

Substituting Equation (2.6) to (2.4) we obtain

$$U_{in} = (ASC_i - \theta_q \alpha_0) + \beta_t t_{in} + \beta_x x_{in} + \theta_q I_{1in} - \theta_q \beta_\delta \delta_{in} - \theta_q \nu_{in} + e_{in}. \tag{2.7}$$

By denoting $\tilde{ASC}_i := ASC_i - \theta_q \alpha_0$, $\theta_\delta := -\theta_q \beta_\delta$ and $\tilde{e}_{in} := -\theta_q \nu_{in} + e_{in}$ we obtain

$$U_{in} = \tilde{ASC}_i + \beta_t t_{in} + \beta_x x_{in} + \theta_q I_{1in} + \theta_\delta \delta_{in} + \tilde{e}_{in}, \tag{2.8}$$

where there is no endogeneity anymore.

The standard IV methods require a variable that satisfies the conditions for an instrument, see e.g. Hausman (1978). MIS requires also conditions for usable indicators: each indicator must be a valid instrument for the system conditional on the other indicator, see e.g. Guevara and Polanco (2016). We assume that I_2 is causing I_1 and that there is no source of unobserved co-variation between them, besides the unobserved attribute q_{in} that causes the endogeneity problem. This assumption renders from Equations (2.2) and (2.3). It is the key assumption of the MIS method, equivalent for the conditions required by the instrumental variables for the application of the control function method.

[3]In linear models, correlation zero between them is required. See Guevara and Polanco (2016) for the formal details.

A limitation of this methodology is that the indicator I_{1in} and the residuals of the regression δ_{in} appear directly in the utility function, as seen in Equation (2.8). This might not be an issue when the purpose of the model is to derive trade offs such as willing to pay estimates or elasticities at the time when the sample was collected, but would be relevant for forecasting. How to overcome this limitation is out of the scope of the thesis, but a research direction would be to write a measurement equation of the indicators that depends on socioeconomic characteristics. By doing this, the indicators could be forecasted and so could be the result of the regression in Equation (2.5). This also applies to the following Sections 2.3.2 and 2.3.3.

2.3.2 MIS method and interactions: first approach

Assume now that the variable q is an interaction term $t_{in} \cdot \xi_n$, where ξ_n is a characteristic of the decision-maker. The specification of the utility function is then

$$U_{in} = \text{ASC}_i + \beta_x x_{in} + \beta_t t_{in} + \beta_\xi t_{in} \xi_n + e_{in}. \tag{2.9}$$

If the term $\beta_\xi t_{in} \xi_n$ is omitted, it would enter the error term. Therefore, the error term would be correlated to t_{in}, causing endogeneity. Suppose again that we have two indicators I_{1in}, I_{2in} for the variable ξ_n, that is, $I_{1in} = \alpha_0 + \alpha_\xi \xi_n + e_{1in}$. If we repeat the derivation from section 2.3.1 we obtain

$$U_{in} = ASC_i + (\beta_t - \theta_\xi \alpha_0) t_{in} + \beta_x x_{in} + \theta_\xi t_{in} I_{1in} - \theta_\xi t_{in} \beta_\delta \delta_{in} + \theta_\xi t_{in} \nu_{in} + e_{in}, \tag{2.10}$$

and by denoting $\tilde{\beta}_t := \beta_t - \theta_\xi \alpha_0$ and $\theta_\delta := -\theta_\xi \beta_\delta$ we obtain

$$U_{in} = ASC_i + \tilde{\beta}_t t_{in} + \beta_x x_{in} + \theta_\xi t_{in} I_{1in} + \theta_\delta t_{in} \delta_{in} + \underbrace{\theta_\xi t_{in} \nu_{in} + e_{in}}_{\tilde{e}_{in}(t)}. \tag{2.11}$$

For this reason, this approach is not further investigated.

2.3.3 MIS method and interactions: correct approach

In order to use the MIS method in the presence of interactions between an attribute t_{in} and an unobserved factor ξ_{in}, we need to assume $t_{in} \cdot I_{1in}$ and $t_{in} \cdot I_{2in}$ to be indicators for $t_{in} \cdot \xi_n$. This assumption might be difficult to fulfill, but it is necessary in order to extend the MIS as proposed below. We define $\xi'_{in} = t_{in} \cdot \xi_n$, $I'_{1in} = t_{in} \cdot I_{1in}$ and $I'_{2in} = t_{in} \cdot I_{2in}$. The following relation can therefore be defined

$$I'_{1in} = \alpha_0 + \alpha_\xi \xi'_{in} + e_{I_{1in}}. \tag{2.12}$$

If the true relation is $I_{1in} = \alpha_0 + \alpha_\xi \xi_n + e_{1in}$, then Equation (2.12) is only an approximation. We can also define

$$I'_{1in} = \gamma_0 + \gamma_1 I'_{2in} + \gamma_t t_{in} + \gamma_x x_{in} + \delta_{in}, \tag{2.13}$$

$$e_{I_{1in}} = \beta_\delta \delta_{in} + \nu_{in}, \tag{2.14}$$

where δ_{in} captures the part of $e_{I_{1in}}$ which is correlated with I'_{1in} and ν_{in} is an exogenous error term.

From Equation (2.12) we obtain $\xi'_{in} = (I'_{1in} - \alpha_0 - e_{I_{1in}})/\alpha_\xi$. By substituting this expression in Equation (2.9), denoting $\theta_\xi = \frac{\beta_\xi}{\alpha_\xi}$; proceeding as in Section 2.3.2, denoting $\tilde{ASC}_i := ASC_i - \theta_\xi \alpha_0$, $\theta_\delta := -\theta_\xi \beta_\delta$ and $\tilde{e}_{in} := -\theta_\xi \nu_{in} + e_{in}$ we obtain

$$U_{in} = \tilde{ASC}_i + \beta_t t_{in} + \beta_x x_{in} + \theta_\xi t_{in} I_{1in} + \theta_\delta \delta_{in} + \tilde{e}_{in}, \tag{2.15}$$

where the endogeneity has been corrected. The model with the MIS correction is estimated in two stages. First δ_{in} is obtained by taking the residual values of Equation (2.13). Second, all parameters of Equation (2.15) are estimated by maximum likelihood. Note that using the full information maximum likelihood would render a one-stage estimation possible.

If a two-stage approach is used, the standard errors from the second stage need to be corrected. Otherwise they will be downward biased. This can be done either by bootstrapping or by considering the analytical formulation. In the control function framework, Petrin and Train (2003) use bootstrapping and Karaca-Mandic and Train (2003) provide the formula for the asymptotic standard errors. Their proposed analytical formula does not apply in our case, as the model specification is different. They show that the results are very similar. The procedure of how to do the bootstrap is explained in detail in Guan (2003).

2.3.4 Integrated choice and latent variable (ICLV) model framework

Instead of using the MIS method to account for the omission of $t_{in}\xi_n$, the ICLV methodology can also be used. The ICLV has been widely addressed in the literature. We refer the interested reader in the theoretical framework to Walker (2001). In Walker and Ben-Akiva (2002) several extensions of random utility models, amongst which ICLV is, are unified in a generalized framework. Finally, Ben-Akiva et al. (2002) discuss the progress and challenges of these models. We assume that the reader is familiar with ICLV and introduce it only briefly.

Let us now consider a model with the same formulation of utility as in Equation (2.9).

The structural equation of the latent variable model is given as follows

$$\xi_n = \eta_0 + \eta s_n + \omega_\xi, \tag{2.16}$$

where η_0, η are (vectors of) parameters to estimate, s_n is a vector of socio-economic characteristics of the respondent n, and ω_ξ is an error term.

The measurement model specifies the following k measurement equations

$$t_{in} I_{kin} = \alpha_k + \lambda_k \xi_n t_{in} + \omega_{I_{kin}}, \tag{2.17}$$

where α_k and λ_k are parameters to estimate, and $\omega_{I_{kin}}$ is a random error term. Note that Equation (2.17) is considered in this way to be consistent with Equation (2.12). To compute the maximum likelihood function, integration over ξ is performed which makes it more computationally complex to estimate. Therefore, the identification of the parameters is not as straight forward as for the MIS method.

2.4 Case Study: Mode choice in Switzerland with RP data

The description of the case study is organized as follows: Section 2.4.1 introduces the dataset that is used, including details of the data collection and some descriptive statistics. It is followed by the model specification in Section 2.4.2. Finally, the results are presented in Section 2.4.3.

2.4.1 Data used: collection and exploratory analysis

The dataset used for the case study was collected in Switzerland between 2009 and 2010 as part of a project to understand mode choice and to enhance combined mobility behavior. It consists of a revealed preferences (RP) survey. Details about the data collection procedure can be found in Bierlaire et al. (2011); Glerum, Atasoy and Bierlaire (2014), and more information about the project can be found in `http://transport.epfl.ch/optima`.

The structure of the questionnaire is as follows. There is a first part consisting of a revealed preferences survey where information on all the trips performed during one day are collected. Respondents report travel time, travel cost, socioeconomic characteristics of themselves and of their household, opinions on a list of statements, mobility habits and what is referred to in Glerum, Atasoy and Bierlaire (2014) as *semi-open questions*. In these semi-open questions, respondents are asked to provide three adjectives to describe each mode. Each observation corresponds to a round trip, not to a single trip. After removing (i) observations where the mode is not reported, (ii) observations corresponding

to respondents who claim to use the car, but answer simultaneously that they do not have access to a car, (iii) those who do not answer to the opinion statement that are used for the modeling and (iv) those who do not report their income level, there is a total of 1,686 observations.

The mode alternatives are public transportation (PT), private motorized modes (PMM) (car, motorbike, etc.) and slow modes (SM) (bike, walk). PMM is also referred to as *Car*. Table 2.1 shows the sample market shares for each of the three considered modes. These are the results after excluding the respondents described above. Of these, only 83 had no access to car. This is taken into account for the modeling. The market shares observed in the sample are coherent with the real market shares in the population (Office fédéral de la statistique, 2012).

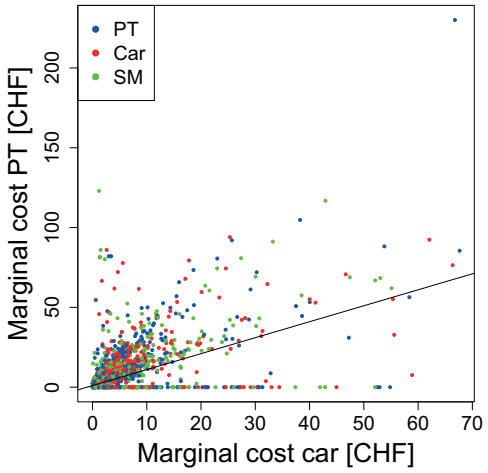
	PT	PMM	SM	Total
Number of observations	456	1,128	102	1,686
Observed market shares (%)	27	67	6	100

Table 2.1: Observed market shares and number of observations for each of the three alternatives in the choice set (public transportation, private motorized modes and slow modes).

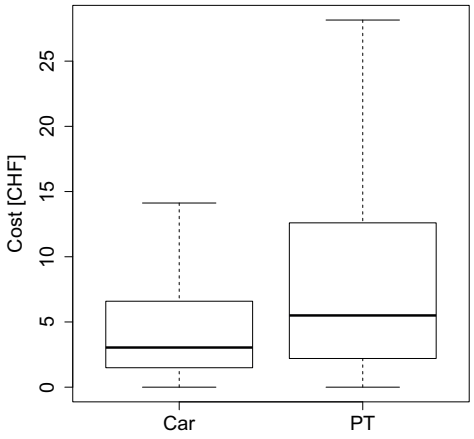
2.4.1.1 Travel time and travel cost

Figure 2.1 shows the travel time and cost both by car and public transportation for each individual. The reported travel time for the chosen mode is not used, instead, it is imputed. Details can be found in Bierlaire et al. (2011).

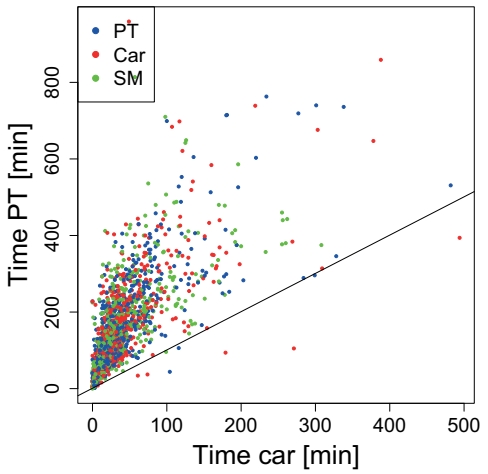
It is observed in Figures 2.1(b),2.1(d) that in general terms car is faster and cheaper than public transportation. This is confirmed by Figure 2.1(c) where we see that there are less than 10 observations where public transportation is faster than car. In Figure 2.1(a), we see that there are several respondents for which the marginal cost by public transportation is zero. This is due to the fact that respondents in the dataset can have several travel cards that makes their marginal cost null. In both figures, the black line represents the $x = y$ line.



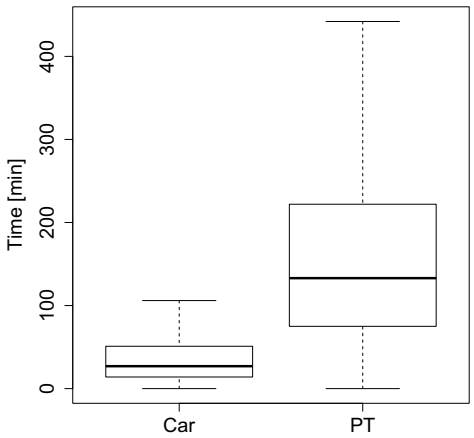
(a) Marginal cost.



(b) Boxplot of marginal cost.



(c) Travel time.



(d) Boxplot of travel time.

Figure 2.1: Plots and boxplots of travel time and travel cost for the different alternatives.

2.4.1.2 Attitudinal questions

Several attitudinal questions related to the car-loving attitude are rated in a 1 to 5 Likert scale by the respondents. The statements that are used in this case study are the following

1. It is difficult to take the public transportation when I travel with my children.
2. With my car I go whenever and wherever.

As described in Section 2.3.3, the indicators that are considered for this case study are the product of these ratings and the travel time. The one corresponding to statement *With my car I go whenever and wherever* is referred to as *flexibility indicator* and the one related to statement *It is difficult to take the public transportation when I travel with my children* is referred to as *convenience indicator*. The correlation between them is 0.88. It is important to note that all the respondents give answers to these indicators. The distribution of the indicators is similar when looking at the responses from the whole sample, and when looking at the individuals who chose to travel by car, public transportation or slow modes separately. Moreover, not all the respondents in the sample have children. Those who do not, respond to Indicator 1 either in a neutral way, or with NA, that is then recoded to value 3 of the Likert indicator. In the reminder of the chapter, the expression *Likert indicator* is used when referring to the 1 to 5 indicators, and the expression *composite indicator* is used to refer to the product of this indicators and travel time.

The assumption that there is no unobserved covariation between the flexibility and the convenience indicators, conditional to the car loving attitude, is a strong one. Both flexibility and convenience are likely to be based on other, yet common, latent psychological constructs. We here assume that their only source of covariance is the car loving attitude. We recognize that such an assumption should be further investigated. Similar assumptions for other unobserved factors are considered in Guevara and Polanco (2016). Further investigation on this is considered future work.

2.4.2 Model specification

Table 2.2 shows the model specification used as the base model for the case study. It is a model with 13 parameters. In the slow modes utility function, only the distance of the trip and the number of bicycles in the household are considered as explanatory variables.

In the public transportation utility, there is the alternative specific constant (ASC), some socioeconomic variables related to the type of neighborhood (rural vs urban) and

Parameter	Public transportation	Car	Slow modes
β_1 (ASC_{PT})	1	0	0
β_2	0	Time car [min]	0
β_3	Travel time by PT [min]	0	0
β_4 (ASC_{car})	0	1	0
β_5	0	Number of children	0
β_6	0	Number of cars	0
β_7	$\frac{\text{Marginal cost of PT}}{\text{Income}}$	$\frac{\text{Marginal cost of car}}{\text{Income}}$	0
β_8	0	Work-related trip	0
β_9	0	French speaking	0
β_{10}	Student	0	0
β_{11}	Urban area	0	0
β_{12}	0	0	dist. [km]
β_{13}	0	0	Number of bicycles

Table 2.2: Base model specification.

to the occupation (student or not), as well as attributes of the mode such as cost and time, where cost is interacted with the income of the respondent. The parameter for time is an alternative specific one, while the parameter related to travel cost is generic for both alternatives.

In the car utility function there is also an ASC and three socioeconomic variables which are if the respondent is from a French speaking part of Switzerland or not, the number of cars in the respondent's household and the number of children in the household. There are also the time and cost of the trip, where the cost is the gasoline cost, and it is again interacted with the income of the respondent. There is also a dummy variable for the trip purpose (if it is work-related or not).

The specifications used for the other two models (MIS and ICLV) are the same except for the parameters associated with each methodology. The base model specification is suspected to suffer from endogeneity issues since it does not consider the interaction between the travel time and the unobserved car lovingness, as discussed earlier in Section 2.3.2.

2.4.3 Results

The presentation of the results is divided in several sections. Sections 2.4.3.1, 2.4.3.2, 2.4.3.3 present the estimation results of the logit, logit with MIS correction and ICLV methodology respectively. They are followed by Sections 2.4.3.4 and 2.4.3.5 where a comparison of the results obtained is performed. All models are estimated using Python-Biogeme, an open source software designed for the estimation of discrete choice models

(Bierlaire, 2016).

2.4.3.1 Base model: Logit

Table 2.3 shows the estimation results for the model specification defined in Table 2.2. The signs are in line with our expectations and the literature. The parameters associated with travel time, travel cost and distance are negative. Moreover, travel time in private modes causes more disutility than travel time in public transportation. This is justified by the fact that the time in public transportation can be used to do other things, while when a person is driving s/he can not do any other activity. Guevara (2017) discusses other potential explanations for this finding.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC (PT)	1.08	0.399	2.71	0.01
2	Travel time [min] (Car)	-0.0272	0.00507	-5.37	0.00
3	Travel time [min] (PT)	-0.00878	0.00169	-5.19	0.00
4	ASC (Car)	0.257	0.440	0.58	0.56
5	No. of children in household (Car)	0.181	0.0699	2.59	0.01
6	Number of cars in household (Car)	1.04	0.125	8.32	0.00
7	$\frac{\text{Marginal cost}}{\text{Income}}$	-0.334	0.0817	-4.08	0.00
8	Work related trip (Car)	-0.659	0.130	-5.06	0.00
9	French speaking (Car)	1.01	0.175	5.79	0.00
10	Student (PT)	2.94	0.481	6.10	0.00
11	Household in urban area (PT)	-0.202	0.134	-1.50	0.13
12	Distance [km] (SM)	-0.204	0.0505	-4.04	0.00
13	No. of bikes in household (SM)	0.390	0.0607	6.43	0.00

Summary statistics
Number of observations = 1686
Number of excluded observations = 579
Number of estimated parameters = 13

$$\begin{aligned} \mathcal{L}(\beta_0) &= -1337.224 \\ \mathcal{L}(\hat{\beta}) &= -880.350 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 913.749 \\ \rho^2 &= 0.342 \\ \bar{\rho}^2 &= 0.332 \end{aligned}$$

Table 2.3: Estimation results for the logit base model.

2.4.3.2 Multiple Indicator Solution method

Table 2.4 shows the estimation results of using the MIS methodology when there is an interaction between travel time and the *car loving* attitude[4]. The approach introduced in

[4]The same model with the roles of the indicators reversed is also estimated. The parameter estimates are comparable in terms of magnitudes and signs, and so are the standard errors, except for the

Section 2.3.3 is used. Bootstrapping is performed to obtain the correct standard errors. All the parameters that appear also in the logit can be interpreted in a similar way, except for travel time by car. The Likert flexibility indicator can take values from 1 to 5, so the travel time parameter is in the range $(-0.0976+1\cdot 0.0172, -0.0976+5\cdot 0.0172) = (-0.0804, -0.0116)$. The β_δ parameter does not have a direct behavioral interpretation, but is derived by the mathematical formulation. It is introduced in Equation (2.14). The fact that parameter 15 is significant, which corresponds to θ_ξ in Equation (2.15), means that there was endogeneity in the logit model.

In order to perform a likelihood ratio test we need to do bootstrapping[5]. Let L be the empirical distribution of the likelihood ratio test statistic. We obtain that

$$\mathrm{P}(L \geq \chi^2_{2,0.05}) = \mathrm{P}(L \geq 5.99) > 0.99.$$

Therefore, the two models are not statistically equivalent, and we conclude that the MIS is preferred.

2.4.3.3 Integrated Choice and Latent Variable method

Finally, an ICLV model is estimated. Results are shown in Table 2.5. Parameters 1-13 can be interpreted as in the case of the logit. In order to understand the rest of the parameters, the structural and measurement equations are introduced. The structural equation for the car-loving attitude is defined as follows:

$$\text{Car loving} = \eta_{Carloving} + \omega, \tag{2.18}$$

where $\omega \sim \mathcal{N}(0, \sigma^2)$ and $\eta_{Carloving}$ is a parameter to estimate. In a classical ICLV approach this structural equation could be more complex. In the case study we consider it as shown in Equation (2.18) so that the results can be compared to those of the MIS method.

The measurement equations are as follows

$$t \cdot I_1 = \alpha_1 + \lambda_1 \cdot t \cdot \text{Car loving} + \omega_1, \tag{2.19}$$

$$t \cdot I_2 = \alpha_2 + \lambda_2 \cdot t \cdot \text{Car loving} + \omega_2, \tag{2.20}$$

where $\omega_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $\omega_2 \sim \mathcal{N}(0, \sigma_2^2)$. For identification reasons, α_1 is normalized to 0, and λ_1 and σ_1 to 1. As explained in Section 2.3.4, Equations (2.19) and (2.20) are

significance of the parameter associated with households in rural areas for which the p-value increases to 0.12.

[5]For each bootstrapped sample we estimate both the MIS and the logit, and calculate the statistic for the given sample. By doing this we obtain the empirical distribution of the LRT statistic and we can compute the probability that this distribution is larger than the $\chi^2_{2,0.05}$ critical value.

Parameter number	Description	Coeff. estimate	Bootstr. std. error	t-stat	p-value
1	ASC (PT)	1.07	0.390	2.75	0.01
2	Travel time [min] (Car)	-0.0976	0.0440	-2.22	0.03
3	Travel time [min] (PT)	-0.00897	0.00197	-4.54	0.00
4	ASC (Car)	0.530	0.484	1.10	0.27
5	No. of children in household (Car)	0.181	0.0735	2.47	0.01
6	Number of cars in household (Car)	0.832	0.200	4.17	0.00
7	$\frac{\text{Marginal cost}}{\text{Income}}$	-0.336	0.104	-3.23	0.00
8	Work related trip (Car)	-0.766	0.142	-5.38	0.00
9	French speaking (Car)	0.953	0.180	5.30	0.00
10	Student (PT)	2.76	0.490	5.54	0.00
11	Household in urban area (PT)	-0.237	0.136	-1.74	0.08
12	Distance [km] (Slow modes)	-0.205	0.0544	-3.77	0.00
13	No. of bikes in household (SM)	0.383	0.0618	6.20	0.00
14	β_δ (Car)	0.536	0.226	2.38	0.02
15	Likert flex. ind. $\times$ travel time [min] (Car)	0.0172	0.105	1.64	0.10

Summary statistics
Number of observations = 1686
Number of excluded observations = 579
Number of estimated parameters = 15

$$\begin{aligned} \mathcal{L}(\beta_0) &= -1337.224 \\ \mathcal{L}(\hat{\beta}) &= -865.351 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 943.747 \\ \rho^2 &= 0.353 \\ \bar{\rho}^2 &= 0.342 \end{aligned}$$

Table 2.4: Estimation results for the MIS method. Standard errors obtained with bootstrapping.

considered to be like this so that the methodology is fully comparable with the MIS. The same model but considering $I_1 = \alpha_1 + \lambda_1 \text{Car loving} + \omega_1$ and $I_2 = \alpha_2 + \lambda_2 \text{Car loving} + \omega_2$ is also estimated, and the results obtained are similar. This suggests that the assumption introduced in Section 2.3.3, that $t_{in} \cdot I_{1in}$ and $t_{in} \cdot I_{2in}$ are indicators for $t_{in} \cdot \xi_n$ is one we can make in this case study. This assumption can also be justified from a behavioral point of view, since the indicators were reported after experiencing the travel time.

Parameter 14, corresponding to the interaction between *Car loving* and travel time, is positive, as expected.

2.4.3.4 Comparison of the methodologies: value of time

Comparison between the models can not be done based on the actual values of the estimators, because the correction of endogeneity introduces a change in scale (Guevara and Ben-Akiva, 2012). We therefore compare the VOT and time elasticities.

In this section the value of time (VOT) estimates are compared across the three methods presented above. The software PythonBiogeme (Bierlaire, 2016) is also used for the

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC (PT)	1.05	0.391	2.69	0.01
2	Travel time [min] (Car)	-0.0680	0.0112	-6.08	0.00
3	Travel time [min] (PT)	-0.00914	0.00173	-5.29	0.00
4	ASC (Car)	0.0870	0.421	0.21	0.84
5	No. of children in household (Car)	0.199	0.0692	2.87	0.00
6	No. of cars in household (Car)	1.09	0.121	9.00	0.00
7	$\frac{\text{Marginal cost}}{\text{Income}}$	-0.346	0.0890	-3.89	0.00
8	Work related trip (Car)	-0.703	0.129	-5.45	0.00
9	French speaking (Car)	0.963	0.171	5.65	0.00
10	Student (PT)	3.38	0.433	7.79	0.00
11	Household in urban area (PT)	-0.216	0.134	-1.62	0.11
12	Distance [km] (SM)	-0.206	0.0500	-4.11	0.00
13	No. of bikes in household (SM)	0.374	0.0598	6.25	0.00
14	Car loving × travel time [min] (Car)	0.0145	0.00303	4.78	0.00
15	$\eta_{Carloving}$	2.68	0.0735	36.42	0.00
16	σ	0.589	0.0176	33.50	0.00
17	α_2	0.000575	0.00766	0.08	0.94
18	λ_2	1.53	0.0453	33.88	0.00
19	σ_2	0.142	0.0189	7.49	0.00

Summary statistics
Number of observations = 1686
Number of excluded observations = 579
Number of estimated parameters = 19

$$\begin{aligned} \mathcal{L}(\beta_0) &= -23121.351 \\ \mathcal{L}(\hat{\beta}) &= -4545.965 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 37150.773 \\ \rho^2 &= 0.803 \\ \bar{\rho}^2 &= 0.803 \end{aligned}$$

Table 2.5: Estimation results for the ICLV method.

simulation of these estimates. It gives as an output the value of the point estimate for each respondent.

Figure 2.2(a) shows a boxplot containing the disaggregate values of VOT of the respondents. We can see that the results obtained with the logit model have a lower spread compared to those of MIS and ICLV, which is expected since the *car loving* attitude is not taken into account. These values have a wider spread than those found by Axhausen et al. (2008). In their research they define four trip purposes: business, commuting, leisure and shopping. Individuals that take the car to go shopping have the lowest VOT, that is of 24.32 CHF/h and individuals that travel for business have the largest one, of 50.23 CHF/h.

Figure 2.2(b) is an alternative representation of the same values, where the VOT have been reordered from the lowest to the highest value. 95% confidence intervals are also represented for each of the methodologies in a lighter color than the mean. We can see that for the logit model we obtain six different values of mean VOT, one for each level of

income. The results obtained for the ICLV are very similar, since the structural equation is given only by the mean plus an error term (see Equation (2.18)). For the mean VOT with the MIS we obtain 30 different values, one per level of income and per answer to the Likert indicator. The higher rate an individual gave to the statement *With my car I can go whenever and wherever*, the lower is his/her VOT. This is in line with what is expected, since a *car lover* is willing to pay less to save a minute of travel time by car compared to a someone with lower affection towards car. The confidence intervals of the MIS are larger than for the logit or the ICLV, but we can also see that some mean VOT obtained with the MIS are outside the confidence intervals obtained by the ICLV and logit. However, the confidence interval bands associated with these VOT are also very large. For this reason we investigate time elasticity in Section 2.4.3.5.

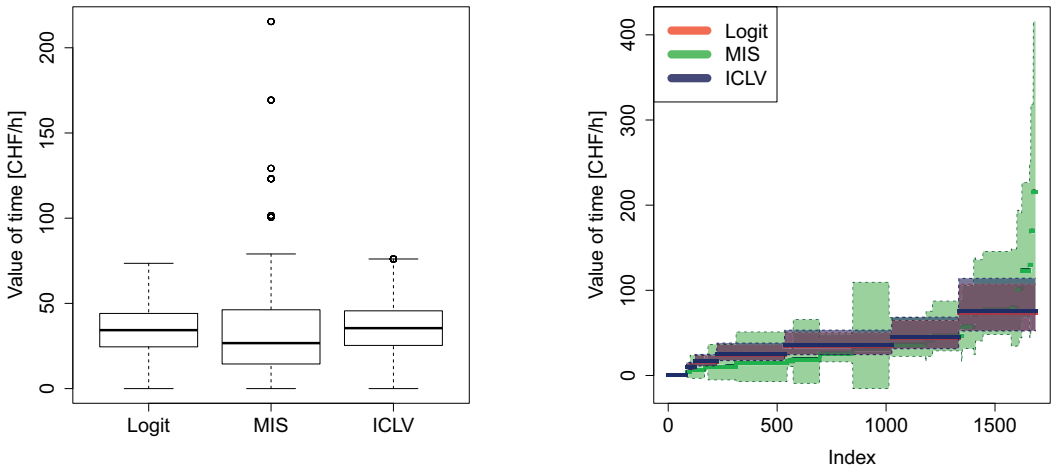
Figure 2.3, is a graphical representation of the VOT for each of the *car loving* and income levels with the MIS method. The value of time for the category of low income and low *car loving* attitude is zero since none of these respondents has access to car. It is interesting to notice that the diagonals of this rectangle have almost the same value of time. For example, an individual with a monthly income of 5,000 CHF that gave the lowest value to the flexibility Likert indicator has the same VOT than a person with a monthly income of 7,000 CHF that rated the indicator with the second value, and the same as an individual with a monthly income of 9,000 CHF that answered with a 3 out of 5 to the flexibility Likert indicator. As expected the highest value of time corresponds to the respondents with the highest income who gave the lowest value to the flexibility Likert indicator. The VOT decreases as income level decreases and as *car lovingness* –represented by the indicator– increases. In this sense, it is interesting to see how a respondent with an income level of at least 15,000 CHF per month has the same VOT as a respondent with a monthly income of 3,250 CHF if the first one rated the indicator with a 5 out of 5, and the second with a 1 out of 5.

2.4.3.5 Comparison of the methodologies: travel time elasticity

The elasticity of travel time represents the percentage of variation in the probability of choosing an alternative following an increase of one percent in the travel time of this alternative.

Table 2.6 shows the weighted average and the 5 and 95 percentiles of travel time elasticity (TE) for both the car and the public transportation alternatives for each of the three methodologies: a logit model, a model with the MIS correction and an ICLV model. Note that to compute the aggregate indicators of demand, the observations have to be weighted to coincide with the real population. Weights calculated by Atasoy et al. (2013) by age, gender and education level using the iterative proportional fitting algorithm.

In all the cases it is negative, as expected, meaning that an increase of travel time in



(a) Boxplot of the VOT for the different methodologies.

(b) Plot of the ordered VOT for the different methodologies with 95% confidence intervals.

Figure 2.2: Representation of the VOT [CHF/h] for car.

a transportation mode decreases the probability of choosing it. It is also observed that the time elasticity for public transportation is larger in absolute value than that of car. This is not what is expected from the parameter estimates. Table 2.3 shows that the parameter related to travel time for public transportation is smaller in absolute value than the parameter related to travel time by car. It becomes clearer by looking at the formula of the elasticity of travel time for an alternative i:

$$E_{t_{in}}^{P_n(i)} = \frac{\partial P_n(i)}{\partial t_{in}} \frac{t_{in}}{P_n(i)}, \tag{2.21}$$

where $P_n(i)$ is the probability of respondent n to choose alternative i with $i \in \{\text{Car}, \text{PT}\}$, and t_{in} is the travel time for respondent n and alternative i. As shown in Figure 2.1, travel time by public transportation is usually longer than by car, so this results in the mean time elasticity for public transportation being larger in absolute value than the mean time elasticity for car. In other words, people are more sensitive to a one minute change in the travel time by car than in the travel time by public transportation. However, they are more sensitive to a 1% change in the travel time by public transportation than to a 1% change in travel time by car.

The results for public transportation do not change much across methods, as expected. For car, the logit model underestimates the mean time elasticity compared to both the MIS and the ICLV. Indeed, a 1% change in travel time by car has an impact of -0.37% on the probability of choosing car, according to the logit model. However, after correcting for endogeneity with either the MIS or the ICLV methodologies, we see that the decrease

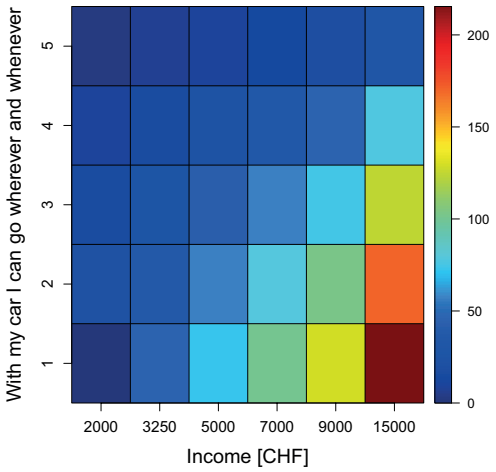


Figure 2.3: Representation of the VOT [CHF/h] for car per income and attitude level using the MIS method.

	Logit			MIS			ICLV		
	5 p.	mean	95 p.	5 p.	mean	95 p.	5 p.	mean	95 p.
Car	-0.52	-0.37	-0.22	-0.88	-0.51	-0.17	-0.60	-0.43	-0.24
PT	-1.29	-0.96	-0.64	-1.39	-0.98	-0.60	-1.31	-0.98	-0.60

Table 2.6: Weighted average, 5 and 95 percentiles of the travel time elasticity for car and public transportation for each of the methodologies used.

would be between 0.43% and 0.51%. Even if the confidence intervals obtained with the MIS are larger than for both the MIS and the ICLV, it is interesting to note that the mean value obtained with the MIS method is in the limit of the 5 percentile for the logit.

The fact that the MIS has larger confidence intervals compared to the other two methods might be due to the two-stage estimation. As future work, it would be interesting to repeat the same in a one-stage estimation and compare the confidence intervals.

It is also interesting to look at the distribution of elasticities across the population, rather than the mean value. Figure 2.4 shows the boxplots across the three different methodologies. Since the spread is very wide – the minimum values are -13.5, -38.8 and -14.4 for the logit, MIS and ICLV values respectively– the boxplot is zoomed in the range $(-1, 0)$. The red cross represents the weighted mean value of TE. We can see that the spread of the boxplot without taking into account the outliers is larger for the ICLV methodology, due to the error terms in the structural and measurement equations. The shape is similar for the MIS and the logit models, but as discussed above, the average

is not, and the tail of the distribution, related to the minimum values, is a lot more negative for the MIS methodology than for the ICLV and the logit, capturing better the extreme values.

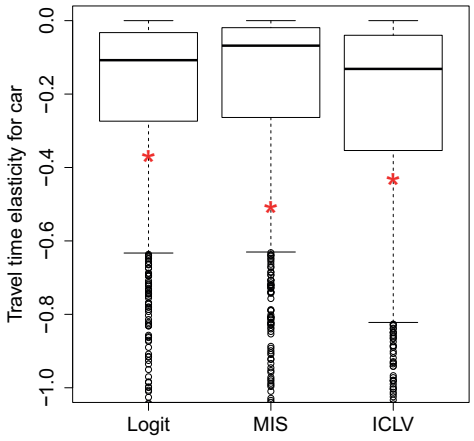


Figure 2.4: Boxplot of the TE for the different methodologies with a red cross representing the mean value.

2.5 Conclusions and future work

We have shown that the Multiple Indicator Solution can also be applied in discrete choice models in the presence of interactions between observed and unobserved attributes in the utility function. Moreover we have tested this methodology with a case study using real data collected in Switzerland. This is the first application of the MIS methodology with revealed preference data. The estimation results obtained are comparable to what is obtained by applying the same correction using the ICLV methodology, and the values of time obtained have larger spread than the results found in the literature since we are taking into account both income and the *car loving* attitude. The distribution of demand indicators such as value of time and time elasticity are also studied. Results reveal that the logit model underestimates the mean travel time elasticity for car compared to both the ICLV and the MIS method. Thanks to the MIS method we can also derive the VOT for different levels of *car lovingness* and income which also reveals interesting results. Moreover, a likelihood ratio test shows that the model with the MIS correction is significantly better than the logit model. In conclusion, the MIS performs as the ICLV or better, and is easier and faster to estimate. The purpose of this case study is to show that the MIS method is operational and that it can be adapted to model interactions between observed and unobserved attributes.

However, the MIS methodology is not free of limitations. An important limitation is that an indicator, as well as the residuals of a regression, appears directly in the utility function. How to do forecasting using this methodology is therefore not trivial. As mentioned, a possibility is to estimate a measurement equation for the unobserved indicators as a function of socioeconomic characteristics of the respondent, and then use these in the utility function.

The difficulty of using the MIS for forecasting might not be a problem if the interest of the application is to compute trade-offs such as VOT estimates, or elasticities at the time when the sample was collected. From a modeling point of view, the MIS method is a logit model with a correction factor. Therefore it has a closed form, and it is computationally a lot faster than the ICLV approach (the estimation time is of less than a second for the MIS method and of around 5 minutes for the ICLV). A potential solution when the model is to be used for forecasting would be to use the MIS approach to identify endogeneity and to find a good model specification, and then apply the ICLV method with the same specification and indicators once it is confirmed. However, as Chorus and Kroesen (2014) point out, ICLV might not be adequate to forecast market shares as a result of a change in the latent variable when the available data is cross sectional. To be able to do so, we need to assume that the causal relationship between the variables and the choice (characterized by the estimated coefficients) is stable over time.

3

Modeling purchases of new cars

This chapter is based on the article:

Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M. (2016). Modeling purchases of new cars: an analysis of the 2014 French market. Accepted for publication in the journal *Theory and decision*

The work has been performed by the candidate under the supervision of prof. Matthieu de Lapparent and prof. Michel Bierlaire.

3.1 Introduction

The automobile sector is of interest for both the public and the private sectors. Governments and other public actors need to understand the car market in order to have valid forecasts of energy consumption, emission levels and even tax revenue. By means of these forecasts they can also derive optimal policy measures to, for instance, promote the use of electric vehicles to reduce emissions.

It is also interesting for private companies. The interest from automobile firms is obvious, but the car market is linked to many other sectors such as those providing the raw materials (steel, chemicals, textiles) and those working with automobiles such as repair and mobility services. Moreover, according to the European Commission, "the EU is

among the world's biggest producers of motor vehicles and the sector represents the largest private investor in research and development (R&D)"[6]

In order to satisfy the needs of these public and private actors it is important to model car ownership, which has many dimensions. Car ownership models can be classified based on several criteria according to de Jong et al. (2004) such as: **i)** the inclusion of supply and demand, **ii)** the aggregation level, **iii)** the time representation (dynamic or static), **iv)** the time horizon (long-term or short-term forecasts), **v)** the inclusion of car-use and other socioeconomic characteristics, and **vi)** the type of market (private or business cars) among others. We focus on the demand side of private cars, in a disaggregate and static framework where we include socioeconomic characteristics of the car buyers. The objective is to have short-term forecasts. This is known as *static disaggregate car-type choice models*.

Our goal is to use revealed preference data to estimate these type of models. This allows to have more realistic demand indicators compared to the ones obtained with stated preference data, such as predicted market shares under several scenarios, willingness to pay and to accept several car attributes and price elasticities. We are particularly interested in the demand for hybrid and electric vehicles. Using revealed preference (RP) data is more challenging. The main difficulties are to define the choice set and to define the attributes of the unchosen alternatives. The main contribution is the way how we define the attributes of the unchosen alternatives. We use the empirical distribution of the attributes and draw from them. The inclusion of supply, which is a major challenge, is not considered in our analysis. The supply is assumed to be exogenous and given. Moreover, we also do not consider the diffusion/adaption process like in Jensen et al. (2017) due to the cross sectional nature of our data.

The remaining of the chapter is structured as follows. Section 3.2 contains a brief literature review, which is followed by the description of the data used in the chapter, and how it is aggregated into different choice alternatives in Section 3.3. In Section 3.4 we discuss the adopted methodological approach, the results of which are discussed in Section 3.5. The application of the model is discussed in Section 3.6. The concluding remarks and future research directions are presented in Section 3.7.

3.2 Literature Review

As mentioned in Section 3.1, we focus on static disaggregate car-type choice models. There are also other types of models that have been used to address electric vehicles' (EV) adoption in the literature, such as agent-based modeling (Adepetu and Keshav, 2017), which are not the focus of this review. The first study dealing with static disag-

[6] http://ec.europa.eu/growth/sectors/automotive

gregate car-type choice models was performed by Lave and Train (1979). For a complete review of the literature on car ownership the reader is referred to de Jong et al. (2004) and more recently to Anowar et al. (2014).

Although it is clear that a choice of a private car is a discrete choice, there does not seem to be consensus in the literature about the definition of the choice set. The two main approaches are defined below. The first approach considers that a car is characterized by its make, model, engine and vintage (Birkeland and Jordal-Jorgensen, 2001). Then, for a given year, there may be over 1,000 alternatives. In this case, sampling of alternatives is usually required for the estimation of the model, although recent developments in model estimation (Mai et al., 2015) allow to estimate large scale multivariate extreme value models.

The second approach prefers an aggregate representation. For example Page et al. (2000) characterize a car by its engine size and fuel type. They have nine alternatives for petrol and seven for diesel. It greatly simplifies the specification and estimation of the model. A similar aggregation of alternatives is used by Hess et al. (2012). This approach is also justifiable from a behavioral point of view, arguing that decision-makers do not explicitly consider large choice sets. Wong et al. (2017) review several methods of aggregating choice alternatives with discrete choice models.

The most popular model in this context is logit (Wu et al., 1999; Choo and Mokhtarian, 2004). However, the Independence from Irrelevant Alternatives (IIA) property of logit, may lead to counterintuitive results when alternatives share unobserved characteristics. It is likely to happen in car-type choice no matter which of the previous two approaches is chosen. Other models have been considered, such as mixtures of logit models, (Brownstone and Train, 1998; McFadden and Train, 2000; Potoglou, 2008), nested logit models (Berkovec and Rust, 1985; McCarthy and Tay, 1998; Mohammadian and Miller, 2002, 2003; Cao et al., 2006) and cross nested logit models (CNL) (Hess et al., 2012).

The interest in electric and hybrid vehicles has risen in the past years, through the analysis of stated preferences data (Glerum, Stankovikj, Thémans and Bierlaire, 2014; Hackbarth and Madlener, 2016; Beck et al., 2013, 2017; Daziano, 2013; Hackbarth and Madlener, 2013; Daziano and Achtnicht, 2014; Brownstone and Train, 1998; Train, 1980; Kim et al., 2014; Rasouli and Timmermans, 2016). Massiani (2014) describes some of the most important limitations of the stated preference surveys being used currently in the literature, and questions the policy recommendations that can be obtained from them.

Studies carried out with revealed preference data are generally quite old, and do not focus on electric and alternative fuel vehicles, mainly because of limitation of these vehicles in revealed preference data. Some examples include Berry et al. (1995, 1998); Train (1986); Berkovec (1985); Train and Winston (2007b); Lave and Train (1979). In these

studies either aggregation of alternatives is used (Train, 1986; Berkovec, 1985; Lave and Train, 1979) or the disaggregate choice set is considered (Berry et al., 1995, 1998; Train and Winston, 2007b). Berkovec and Rust (1985) instead use sampling of alternatives, where 14 out of 785 alternatives are sampled for estimation. The main difficulty when using revealed preference data is to impute the attributes of the unchosen alternatives, in particular if there is an aggregation of alternatives. The studies cited above address this by imputing mean values of the attributes for each unchosen alternative.

We fill the gap in the literature of vehicle choice by proposing an alternative way of imputing the attributes of the unchosen alternatives, based on multiple imputations using the empirical distributions of the attributes for each alternative. A similar approach has been used in the context of residential location choice by Li (2014), when merging two datasets with different level of aggregation. To the best of our knowledge, it is also the first study of car-type choice to focus on electric and hybrid vehicles in the context of the whole market using revealed preference data. Due to the lack of variability in the autonomy (the range) of electric vehicles that are currently in the market, we need to take the value of the willingness to pay for range from the literature. By doing this we are able to forecast the impact of an increase of the range on the market shares of electric vehicles. We validate our model with demand indicators such as market shares, elasticities and willingness to pay and to accept.

3.3 Data and data aggregation

We use a dataset reporting sales of new cars in France in 2014. Each observation corresponds to the purchase of a new car. The dataset reports over 40000 purchases. However, after selecting the variables that we use in the model and removing the missing values for any of them, 18804 observations are used for this study. It is left for future work to recover some of these missing values.

In our approach we decide to consider a car-type as a combination between a market segment and a fuel type. The market segments are *full*, *luxury*, *medium*, *multipurpose vehicles (MPV)*, *off-road* and *small*[7]. The fuel types considered are *diesel*, *petrol*, *electric* and *hybrid*. Table 3.1 gives some examples of cars belonging to each market segment.

Out of the 18804 observations, only 657 report the purchase of a hybrid vehicle. Moreover, they are always combined with either diesel or petrol. Consequently, we consider *hybrid* as a market segment rather than as a fuel type. Therefore, we have a total of 15 alternatives summarized in Table 3.2, together with the number of observations corresponding to each alternative after removing any missing values. Note that we should have 21 alternatives (3 fuel types multiplied by 7 market segments). The six missing

[7]These segments are derived from the European Commission's segmentation

Market segment	Car
Full	Ford Taurus Toyota Avalon Hyundai Grandeur
Luxury	Mercedes-Benz S-Class Audi A8 BMW 7 Series
Medium	Opel Astra Honda Civic Audi A3
Multipurpose vehicle	Renault Espace Volkswagen Sharan Mercedes-Benz Vito
Off-road	Land Rover Freelander Chevrolet Captiva BMW X5
Small	Opel Corsa Ford Fiesta Toyota Yaris

Table 3.1: Examples of cars that belong to each market segment.

alternatives are the combinations of *electric* vehicles with any market segment except *small*. This is because these alternatives do not exist in the French market. The only exception is *electric luxury* vehicles, that do exist (e.g: Tesla), but represent a negligible part of the car market. For this reason, we decide to remove this alternative (*electric luxury*) from the analysis.

From this definition of alternatives, it is obvious that alternatives that share either fuel type or market segment share unobserved attributes. Figure 3.1 proposes a correlation structure derived from the multi-dimensional nature of the choice set presented in Table 3.2. This correlation structure is used in the cross nested logit model presented in Section 3.4.

We assume that our dataset is representative of the population of new car buyers from an exogenous point of view (i.e: from a socioeconomic point of view). It is important to note that it might not be representative of the whole French population, but this is not an issue. For the representativeness of the choices, we are able to replicate the real market shares by applying the correction of the alternative specific constants as described by Train (2009).

Alternative	Market segment	Fuel type	Nbr. of obs.
1	Full	Diesel	323
2	Luxury	Diesel	178
3	Medium	Diesel	2226
4	MPV	Diesel	1375
5	Off-road	Diesel	2044
6	Small	Diesel	5538
7	Hybrid	Diesel	161
8	Full	Petrol	68
9	Luxury	Petrol	54
10	Medium	Petrol	663
11	MPV	Petrol	310
12	Off-road	Petrol	265
13	Small	Petrol	5037
14	Hybrid	Petrol	496
15	Small	Electric	66
Total			18804

Table 3.2: List of alternatives in the choice set and number of observations (after removing missing values).

3.4 Methodological approach and model specification

This section contains the methodological approach used to estimate the choice model. We define the choice model used in Section 3.4.1 followed by the model specification, the nesting structure and the definition of the variables used in Section 3.4.2. Finally, we describe how we import the parameter associated with range from the literature in Section 3.4.3.

3.4.1 Choice model

The choice models used in our study are both a logit model and a cross nested logit. We consider the logit model to be the benchmark in order to show the importance of accounting for the correlation across alternatives. We use the cross nested model in the application of the model.

The cross nested logit model allows to overcome the IIA property. Suppose that the choice set $\mathcal{C}$ is formed by M nests, $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_M$. The parameters α_{im} represent the degree of membership of alternative i to nest $\mathcal{C}_m$. For identification purposes it is bounded between 0 and 1 and the $\sum_{m=1}^{M} \alpha_{im} = 1, \forall i$. The expression of the choice

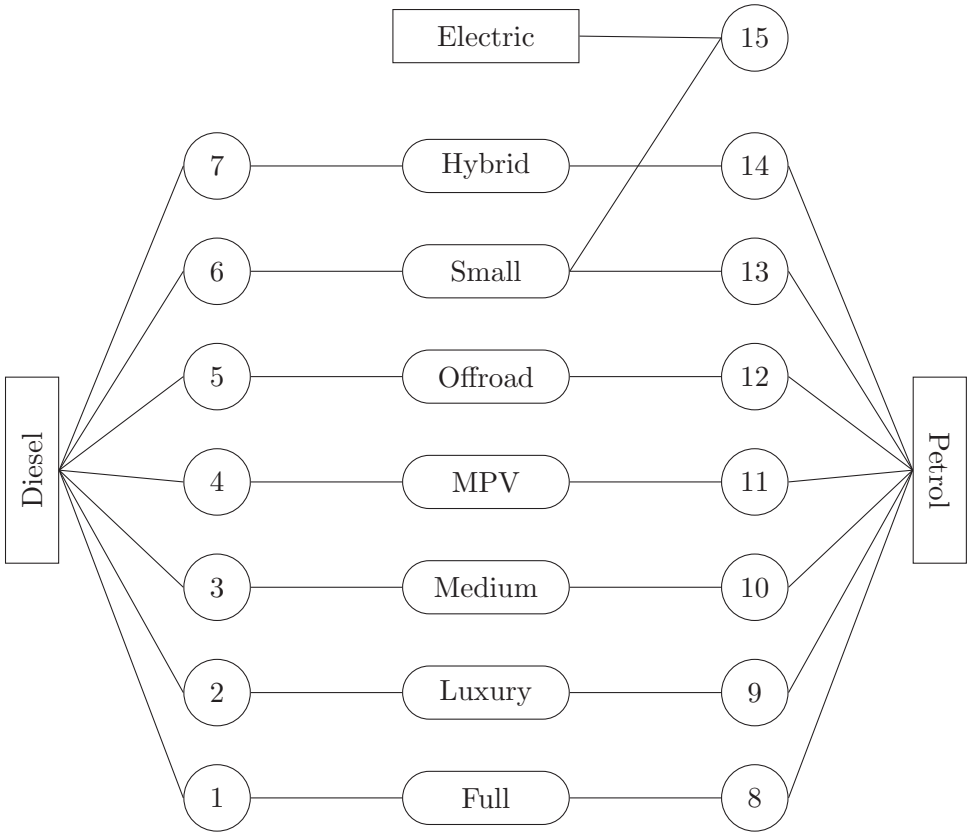


Figure 3.1: Cross nested structure.

probabilities for a cross nested logit model are:

$$\mathrm{P}(i \mid \mathcal{C}_n) = \sum_{m=1}^{M} \frac{\left(\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\frac{\mu_m}{\mu}} \mathrm{e}^{\mu_m V_{jn}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j \in \mathcal{C}_n} \alpha_{jp}^{\frac{\mu_p}{\mu}} \mathrm{e}^{\mu_p V_{jn}}\right)^{\frac{\mu}{\mu_p}}} \cdot \frac{\alpha_{im}^{\frac{\mu_m}{\mu}} \mathrm{e}^{\mu_m V_{in}}}{\sum_{j \in \mathcal{C}_n} \alpha_{jm}^{\frac{\mu_m}{\mu}} \mathrm{e}^{\mu_m V_{jn}}}, \tag{3.1}$$

where $\mathcal{C}_n$ is the choice set of individual n, μ is the scale parameter of the model, normalized to 1, and μ_m, $m = 1, ..., M$ are the scale parameters of each nest, estimated from the data. V_{in} is the deterministic part of the utility function for individual n and alternative i. For details on the normalization of the μ parameters, and a more detailed analysis of the cross nested logit model, the reader is referred to Bierlaire (2006) and Abbe et al. (2007).

3.4.2 Definition of variables and model specification

We consider a logit and a cross nested logit model, with a linear in parameter specification for the utility functions. Table 3.3 shows the variables considered in the model.

They are divided in (i) attributes of the recently purchased car and (ii) socioeconomic characteristics of the main driver of the car and/or her household. It is important to note that the car attributes are those reported by the individuals, and not catalog attributes.

	Variable	Definition
Attributes	price	Purchase price after discounts and government schemes [€/10]
	cons	Fuel consumption [l/10km]
	max_power	Engine power [brake horsepower (bhp)]
	range_EV	Reported average range achieved from a full charge [km]
Socioeconomic characteristics	agglomeration	1 if main driver lives either in a city or in the suburbs
	town_rural	1 if main driver lives either in a town, village or rural area
	university	1 if education level of main driver is at least a bachelor degree
	nbr. cars	Total number of cars in regular use in the household
	nbr. adults	Number of adults in the household (including main driver)
	nbr. child.	Number of children in the household (aged 18 or less)
	income	10000 if annual gross household income ≤ 10000 [€]
		15000 if annual gross household income $\in [10000, 20000)$ [€]
		25000 if annual gross household income $\in [20000, 30000)$ [€]
		35000 if annual gross household income $\in [30000, 40000)$ [€]
		45000 if annual gross household income $\in [40000, 50000)$ [€]
		55000 if annual gross household income $\in [50000, 60000)$ [€]
		65000 if annual gross household income $\in [60000, 75000)$ [€]
		87500 if annual gross household income $\in [75000, 100000)$ [€]
		112500 if annual gross household income $\in [100000, 125000)$ [€]
		150000 if annual gross household income $\in [125000, 175000)$ [€]
		200000 if annual gross household income ≥ 175000 [€]

Table 3.3: Definition of the variables used in the model.

Since the dataset consists of revealed preference choices, we have no direct access to the attributes of the unchosen alternatives and they have to be imputed. The state-of-the-art is to impute the attribute of an unchosen alternative as the mean of that attribute from the chosen alternatives (Train, 1986; Berkovec, 1985). In other words, if an individual chose a *small petrol* car, the *max_power* of the *off road petrol* car is usually imputed as the mean *max_power* of the observed *off road petrol* cars. Instead, we perform multiple imputations (see, for example, Schafer (2000)), by considering the empirical distribution of each attribute for a given alternative. This distribution consists in the observed values of other people's chosen alternatives. Algorithm 1 shows how, where K is the number of multiple imputations, N is the set of respondents, $\mathcal{C}_n$ is the set of alternatives available to individual n (as discussed in Section 3.3) and Y is the set of cars[8]. We define $t : Y \to \mathcal{C}_n$ as the function that maps each car with its car-type such that $t(y) = i$ if car y belongs to alternative (car-type) i. Note that $t(\cdot)$ is surjective but not injective. That is, each

[8]Here, a *car* is defined by a combination of make-model-type. The alternatives are car-types, as defined in Table 3.2.

car belongs to a car-type, and two different cars can belong to the same car-type. We estimate the model repeatedly with the different datasets $D_1, ..., D_k$ built as defined by Algorithm 1. We denote by $\widehat{\theta}^k$ the maximum likelihood estimates of the parameters obtained using dataset D_k. Therefore, we obtain a distribution of the model parameters rather than a point estimate.

Algorithm 1: Attributes of all alternatives.

Data: number of multiple imputations K, set of respondents N, set of alternatives $\mathcal{C}_n$, set of cars Y, vector of attributes for each car x_y

Result: Datasets $D_1, ..., D_k$ containing attributes of chosen and unchosen alternatives

```
1  begin
2      for k = 1 : K do
3          for n ∈ N do
4              for i ∈ C_n do
5                  if individual n chose alternative i then
6                      attributes of alternative i ← attributes of chosen car
7                  else
8                      select randomly (with equal probability) a car y such that t(y) = i
9                      attributes of alternative i ← attributes of car y
10         D_k ← attributes of chosen and unchosen alternatives
```

Tables 3.4 and 3.5 show the model specification. Note that both price and fuel consumption are interacted with income. The fuel consumption is also multiplied by the mean fuel price (diesel or petrol), calculated for 2014 in France (Institut national de la statistique et des études économiques, 2016c). Petrol price is denoted as *pp* and diesel price is denoted as *pd* in the table. The rest of the variables appear linearly in the model. Note that some of them are rescaled for numerical reasons. Note also that the specification of the utility functions is the same for the logit, and the cross nested logit models.

Parameter	1	2	3	4	5	6	7	8
ASC_i	1 in the utility function of alternative $i \in (2, 15)$, 0 in the rest							
$\beta_{\text{price\_inc}_i}$	$\frac{\text{price}_i \cdot 100}{\text{income}}$ in the utility function of alternative i, 0 in the rest							
$\beta_{\text{inc\_full}}$	$\frac{\text{income}}{10000}$	0	0	0	0	0	0	$\frac{\text{income}}{10000}$
$\beta_{\text{inc\_luxury}}$	0	$\frac{\text{income}}{10000}$	0	0	0	0	0	0
$\beta_{\text{inc\_medium}}$	0	0	$\frac{\text{income}}{10000}$	0	0	0	0	0
$\beta_{\text{inc\_MPV}}$	0	0	0	$\frac{\text{income}}{10000}$	0	0	0	0
$\beta_{\text{inc\_offroad}}$	0	0	0	0	$\frac{\text{income}}{10000}$	0	0	0
$\beta_{\text{inc\_hybrid}}$	0	0	0	0	0	0	$\frac{\text{income}}{10000}$	0
$\beta_{\text{nbr\_adults\_small}}$	0	0	0	0	0	nbr. adults	0	0
$\beta_{\text{nbr\_chilren\_small}}$	0	0	0	0	0	nbr. child.	0	0
$\beta_{\text{nbr\_cars\_lux}}$	0	nbr. cars	0	0	0	0	0	0
$\beta_{\text{nbr\_cars\_hybrid}}$	0	0	0	0	0	0	nbr. cars	0
$\beta_{\text{university}}$	0	0	0	0	0	0	university	0
$\beta_{\text{town\_rural\_EV}}$	0	0	0	0	0	0	0	0
$\beta_{\text{town\_rural\_hybrid}}$	0	0	0	0	0	0	town_rural	0
$\beta_{\text{conso\_inc}}$	$\frac{\text{cons}_1 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_2 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_3 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_4 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_5 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_6 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_7 \cdot \text{pd} \cdot 100}{\text{income}}$	$\frac{\text{cons}_8 \cdot \text{pp} \cdot 100}{\text{income}}$
$\beta_{\text{max\_power}}$	$\frac{\text{max\_power}_1}{10}$	$\frac{\text{max\_power}_2}{10}$	$\frac{\text{max\_power}_3}{10}$	$\frac{\text{max\_power}_4}{10}$	$\frac{\text{max\_power}_5}{10}$	$\frac{\text{max\_power}_6}{10}$	$\frac{\text{max\_power}_7}{10}$	$\frac{\text{max\_power}_8}{10}$
$\beta_{\text{range\_EV}}$	0	0	0	0	0	0	0	0

Table 3.4: Model specification (part 1/2).

Parameter	9	10	11	12	13	14	15
ASC_i	1 in the utility function of alternative $i \in (2, 15)$, 0 in the rest						
$\beta_{\text{price\_inc}_i}$	$\frac{\text{price}_i \cdot 100}{\text{income}}$ in the utility function of alternative i, 0 in the rest						
$\beta_{\text{inc\_full}}$	0	0	0	0	0	0	0
$\beta_{\text{inc\_luxury}}$	$\frac{\text{income}}{10000}$	0	0	0	0	0	0
$\beta_{\text{inc\_medium}}$	0	$\frac{\text{income}}{10000}$	0	0	0	0	0
$\beta_{\text{inc\_MPV}}$	0	0	$\frac{\text{income}}{10000}$	0	0	0	0
$\beta_{\text{inc\_offroad}}$	0	0	0	$\frac{\text{income}}{10000}$	0	0	0
$\beta_{\text{inc\_hybrid}}$	0	0	0	0	0	$\frac{\text{income}}{10000}$	0
$\beta_{\text{nbr\_adults\_small}}$	0	0	0	0	nbr. adults	0	0
$\beta_{\text{nbr\_chilren\_small}}$	0	0	0	0	nbr. child.	0	0
$\beta_{\text{nbr\_cars\_lux}}$	nbr. cars	0	0	0	0	0	0
$\beta_{\text{nbr\_cars\_hybrid}}$	0	0	0	0	0	nbr. cars	0
$\beta_{\text{university}}$	0	0	0	0	0	university	university
$\beta_{\text{town\_rural\_EV}}$	0	0	0	0	0	0	town_rural
$\beta_{\text{town\_rural\_hybrid}}$	0	0	0	0	0	town_rural	0
$\beta_{\text{conso\_inc}}$	$\frac{\text{cons}_9 \cdot \text{pp} \cdot 100}{\text{income}}$	$\frac{\text{cons}_{10} \cdot \text{pp} \cdot 100}{\text{income}}$	$\frac{\text{cons}_{11} \cdot \text{pp} \cdot 100}{\text{income}}$	$\frac{\text{cons}_{12} \cdot \text{pp} \cdot 100}{\text{income}}$	$\frac{\text{cons}_{13} \cdot \text{pp} \cdot 100}{\text{income}}$	$\frac{\text{cons}_{14} \cdot \text{pp} \cdot 100}{\text{income}}$	0
$\beta_{\text{max\_power}}$	$\frac{\text{max\_power}_9}{10}$	$\frac{\text{max\_power}_{10}}{10}$	$\frac{\text{max\_power}_{11}}{10}$	$\frac{\text{max\_power}_{12}}{10}$	$\frac{\text{max\_power}_{13}}{10}$	$\frac{\text{max\_power}_{14}}{10}$	$\frac{\text{max\_power}_{15}}{10}$
$\beta_{\text{range\_EV}}$	0	0	0	0	0	0	$\frac{\text{range\_EV}}{100}$

Table 3.5: Model specification (part 2/2).

The choice set $\mathcal{C}_n$ of each individual n is considered to be the universal choice set $\mathcal{C}$, containing the 15 alternatives defined in Figure 3.1. There is some discussion in the literature about whether this is a behaviorally correct assumption. Frejinger and Bierlaire (2010) discuss about the consideration set generation algorithms in the context of route choice, and show that in many cases, they do not contain the chosen alternative. They state that the bias of not including a considered alternative in the choice set is larger than including non chosen alternatives in it. Moreover, in our case, we only have 15 alternatives in the universal choice set, making it therefore operational to estimate the choice model using it.

Nesting structure The nesting structure is defined as in Figure 3.1, where the numbered circles represent the 15 alternatives, the oval shape boxes represent the nest related to the market segment, and the rectangle boxes represent the nests related to the fuel type. We define one membership parameter, α_{MS}, that defines the membership to the *market segment* nests. Then, $1 - \alpha_{MS}$ gives the membership to the *fuel type* nests. More general specifications were tested, but the resulting models were not identified. This is a strong assumption, and more investigation is left for future research. We define also four scale parameters, μ_k, $k \in \{\text{offroad, small, diesel, petrol}\}$. μ_{electric} has to be normalized to one, because it only contains one alternative. μ_ℓ, $\ell \in \{\text{hybrid, MPV, luxury, full, medium}\}$ are normalized to one because they reach the lower bound when we try to estimate them.

3.4.3 Parameter associated with range for electric vehicles

Due to the lack of variability in the range of electric vehicles in the data, the parameter $\beta_{\text{range\_EV}}$ cannot be estimated with sufficient precision. Since the willingness to pay for range is well studied in the literature based on stated preference data, and is known to be one of the determinants of electrical vehicle purchase, we import it from the literature and use it in our model. Dimitropoulos et al. (2013) perform a meta-analysis based on 129 willingness to pay estimates and find that consumers are willing to pay between 66 and 75US$ on average for a 1-mile increase in range, which is equivalent to between 30.8 and 35.0€/km[9]. For the results shown in Sections 3.5 and 3.6, we consider the value 34€/km. We note $\text{WTP}(\text{range}_{lit}) = 34$.

From the definition of willingness to pay for range (since alternative 15 is the one related

[9]For the change in units, we consider the mean exchange rate between US$ and € for 2014 which is 1.33$/€ according to the European Central Bank.

to electric vehicles) we obtain:

$$\text{WTP}(\text{range}_{15,n}) = -\frac{\frac{\partial V_{15,n}}{\partial range_{15,n}}}{\frac{\partial V_{15,n}}{\partial price_{in}}} = -\frac{\beta_{range\_EV} \cdot \text{income}_n}{1000 \cdot \beta_{price\_inc\_15}}, \tag{3.2}$$

and by equalizing $\text{WTP}(\text{range}_{lit}) = \text{WTP}(\text{range}_{15,n})$:

$$\beta_{\text{range\_EV}} = -34 \cdot \frac{1000 \cdot \beta_{price\_inc\_15}}{\text{income}_n}. \tag{3.3}$$

We define $\beta_{\text{range\_EV}}$ as defined in Equation (3.3) and estimate all the parameters simultaneously.

3.5 Results

The estimation results for both the logit and the cross nested logit models are reported in Table 3.6. The reported parameters are the means of the parameters obtained with the $K = 50$ realizations of the multiple imputations method. The standard errors of each parameter s_p are obtained as follows (see Schafer (2000) for more details). Let $\widehat{\theta}_p^k$ be the value of the estimated parameter at imputation k, and $\bar{\widehat{\theta}}_p = \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_p^k$. Let U_p^k be the variance associated with parameter $\widehat{\theta}_p^k$ at imputation k. Then the total variance associated with $\bar{\widehat{\theta}}_p$ has two components, the within-imputation variance (U_p), and the between-imputation variance (B_p). They are calculated as follows

$$U_p = \frac{1}{K}\sum_{k=1}^{K} U_p^k, \tag{3.4}$$

$$B_p = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\theta}_p^k - \bar{\widehat{\theta}}_p)^2. \tag{3.5}$$

The total variance is then

$$s_p^2 = U_p + (1 + K^{-1})B_p. \tag{3.6}$$

	Logit		CNL	
	mean param.	t-test[10]	mean param.	t-test[10]
ASC_2	-1.89	-3.56	-1.84	-3.63
ASC_3	1.88	6.13	1.76	5.55
ASC_4	2.44	8.13	2.19	6.50

[10]The reported t-tests are against zero for all parameters except for the μ parameters. For the μ parameters, the reported t-tests are against one.

ASC_5	1.94	6.51	1.85	5.83
ASC_6	3.64	12.2	3.04	8.38
ASC_7	0.210	0.510	0.273	0.674
ASC_8	-2.38	-8.64	-2.29	-8.12
ASC_9	-3.09	-5.43	-3.08	-5.66
ASC_{10}	0.469	1.44	0.378	1.08
ASC_{11}	1.20	3.95	0.990	2.76
ASC_{12}	-0.277	-0.894	0.271	0.563
ASC_{13}	3.45	11.5	2.86	7.69
ASC_{14}	1.23	3.49	1.10	2.80
ASC_{15}	0.0996	0.165	0.0298	0.0441
$\beta_{\text{inc\_full}}$	0.143	8.28	0.120	6.68
$\beta_{\text{inc\_luxury}}$	0.203	9.81	0.178	8.53
$\beta_{\text{inc\_medium}}$	0.0604	5.67	0.0390	3.68
$\beta_{\text{inc\_MPV}}$	0.0154	1.51	0.00322	0.340
$\beta_{\text{inc\_offroad}}$	0.117	13.2	0.0744	5.50
$\beta_{\text{inc\_hybrid}}$	0.0938	6.49	0.0640	4.35
$\beta_{\text{nbr\_adults\_small}}$	-0.0914	-3.44	-0.0697	-3.24
$\beta_{\text{nbr\_chilren\_small}}$	-0.235	-13.5	-0.192	-8.66
$\beta_{\text{nbr\_cars\_lux}}$	0.292	2.90	0.291	3.08
$\beta_{\text{nbr\_cars\_hybrid}}$	-0.260	-3.75	-0.243	-3.69
$\beta_{\text{university}}$	0.179	2.06	0.169	1.97
$\beta_{\text{town\_rural\_EV}}$	0.555	1.61	0.536	1.64
$\beta_{\text{town\_rural\_hybrid}}$	-0.270	-2.90	-0.230	-2.47
$\beta_{\text{price\_inc\_1}}$	-0.128	-5.96	-0.122	-5.79
$\beta_{\text{price\_inc\_2}}$	-0.102	-4.38	-0.0981	-4.35
$\beta_{\text{price\_inc\_3}}$	-0.134	-12.8	-0.121	-10.68
$\beta_{\text{price\_inc\_4}}$	-0.109	-13.3	-0.0913	-9.81
$\beta_{\text{price\_inc\_5}}$	-0.116	-14.8	-0.0997	-12.01
$\beta_{\text{price\_inc\_6}}$	-0.107	-16.8	-0.0832	-11.47
$\beta_{\text{price\_inc\_7}}$	-0.169	-5.97	-0.170	-5.89
$\beta_{\text{price\_inc\_8}}$	-0.0716	-3.10	-0.0703	-3.08
$\beta_{\text{price\_inc\_9}}$	-0.136	-5.06	-0.124	-5.09
$\beta_{\text{price\_inc\_10}}$	-0.140	-7.55	-0.131	-7.06
$\beta_{\text{price\_inc\_11}}$	-0.112	-8.43	-0.0947	-6.82
$\beta_{\text{price\_inc\_12}}$	-0.0996	-7.14	-0.0991	-7.36
$\beta_{\text{price\_inc\_13}}$	-0.0943	-11.3	-0.0720	-8.93
$\beta_{\text{price\_inc\_14}}$	-0.146	-10.3	-0.132	-9.02
$\beta_{\text{price\_inc\_15}}$	-0.531	-4.22	-0.461	-4.13
$\beta_{\text{conso\_inc}}$	-0.105	-4.90	-0.0774	-4.19

$\beta_{\text{max\_power}}$	0.0567	16.4	0.0481	12.8
μ_{offroad}	-	-	1.34	1.35
μ_{small}	-	-	1.77	2.05
μ_{diesel}	-	-	4.25	1.55
μ_{petrol}	-	-	2.92	0.702
α_{MS}	-	-	0.610	7.04

Table 3.6: Mean of the parameter estimates. Number of multiple imputations: K=50.

Unless pointed out, the following interpretations are valid for both the logit and the cross nested logit models.

Income The interactions between the income level and the market segment have the expected relative magnitudes. The normalized market segment is *small*. Therefore, the interpretation of the results is that people with larger income levels have a larger preference towards *luxury* vehicles, then *full*, followed by *offroad* and *hybrid*, with almost the same magnitude. The less preferred alternatives, all else being equal, for people with larger income are *medium*, then *MPV* and the less preferred is the reference level *small*. Note that $\beta_{\text{inc\_MPV}}$ is not significantly different from zero, so there is no significant difference between the preference towards *MPV* and *small*.

Other socioeconomic characteristics We also model the effects of number of children, of adults and of vehicles in a household, the education level and the residence location.

From the negative values of $\beta_{\text{nbr\_adults\_small}}$ and $\beta_{\text{nbr\_children\_small}}$ we can conclude that the more people live in a household (either adults or children), less likely it is no have a *small* vehicle compared to households with less people. Moreover, the number of children has a stronger effect in the decrease of the probability of buying a *small* vehicle than the number of adults.

From the estimation results we can also conclude that the larger the number of cars in a household, the more likely it is to buy a luxury car (since $\beta_{nbr\_cars\_lux} > 0$). Similarly, the larger the number of cars in a household, the less likely it is a hybrid one ($\beta_{nbr\_cars\_hybrid} < 0$). From the positive value of $\beta_{\text{university}}$ we can conclude that individuals who go to university are more likely to buy hybrid and pure electric vehicles compared to people that do not go to university. For the residence location, we find a surprising result: individuals living in towns or rural areas are more likely to buy an EV than those living in a city or in the suburbs ($\beta_{\text{town\_rural\_EV}} > 0$). For hybrid cars it is however the opposite: individuals living in cities and suburbs are more likely to buy one than those living in towns or rural areas ($\beta_{\text{town\_rural\_hybrid}} < 0$).

Price interacted with income Both pairwise t-test comparisons between the parameter estimates, and a likelihood ratio test reject the hypothesis of generic price parameters. All the price parameters are negative, as expected, and individuals are more sensitive to high prices for electric vehicles (alternative 15) than to any other alternative (since $\beta_{\text{price\_inc\_15}}$ is the largest parameter in absolute value).

By estimating these parameters, we assume that there is presence of income effect. Since they are significant, we consider that this hypothesis is verified. Jara-Daz and Videla (1989) show how to test for income effect by considering only the cost variable that enters the utility function in a quadratic form. Moreover, since income appears both interacted with price, and linearly in the utility function, its effect is not immediate to interpret from these values. More analysis would be needed to discuss this further.

Other attributes of the alternatives The fuel consumption is multiplied by 1.48 €/ℓ, that is the mean petrol price in France in 2014 for the petrol alternatives, and by 1.29 €/ℓ, that is the mean diesel price in France in 2014 for the diesel alternatives (Institut national de la statistique et des études économiques, 2016c). The variable is therefore a proxy to the running costs. We consider the interaction between fuel consumption and household income (see Table 3.4) analogously as we do for price (or the fixed cost). As expected, $\beta_{\text{conso\_inc}}$ is negative, meaning that all else being equal, individuals prefer cars with less fuel consumption. We also model the engine power, that has a positive effect. All else equal, individuals prefer vehicles with more power, as expected.

Nest and membership parameters Three of the four reported nest parameters are not significantly different from one at the 5% level. However, the proposed nesting structure makes behavioral sense, and a likelihood ratio test shows that the cross nested logit model is preferred to the logit model. This leads us to keep the cross nested logit, meaning that the alternatives that belong to the same nest share unobserved characteristics. The other six nest parameters are to be fixed to one. μ_{electric} is fixed to one because it only contains one alternative, so it cannot be identified. The other five (μ_{full}, μ_{luxury}, μ_{MPV}, μ_{hybrid}, μ_{medium}) are fixed to one because when we try to estimate them they reach the lower boundary 1.

The membership parameter α_{MS} is between 0 and 1 and it represents how much (out of one) an alternative is explained by the market segment. The fact that it is larger than 0.5 means that an alternative *belongs more* to its corresponding market segment than to its corresponding fuel type. Note that we consider the same membership parameter to market segment and to fuel type for all the alternatives.

3.6 Application of the model

In this section we apply the model described above in order to obtain demand indicators such as price elasticities (Section 3.6.1), aggregate market shares under different policy scenarios (Section 3.6.2) and willingness to pay and to accept (Section 3.6.3). We consider only the cross nested logit.

For the application of the model, instead of doing multiple imputations (as we do in the estimation process) we impute each attribute of an unchosen alternative by the mean value of each attribute for the chosen alternatives[11]. In other words, for an individual n that chose alternative i, the values of an attribute of the unchosen alternative j, x_{jn} are imputed as the average of attribute x of those individuals who chose alternative j, $\bar{x}_j$. Moreover, in order to replicate the population market shares in the base case, we need to calibrate the alternative specific constants as described by Train (2009, p. 67).

3.6.1 Price elasticities

Let p_{in} be the current value of the price variable, and $p_{jn}^{+} = p_{jn} + \Delta p_{jn}$ the future value. Keeping all other variables at their current values, we denote $P_n(i)$ the choice probability of alternative i and $P_n^{+}(i) = P_n(i) + \Delta P_n(i)$ the choice probability involving p_{jn}^{+}. The disaggregate arc elasticity for individual n is defined as follows:

$$E_{\Delta p_{jn}}^{\Delta P_n(i)} = \frac{\Delta P_n(i)}{P_n(i)} \cdot \frac{p_{jn}}{\Delta p_{jn}}, \quad \forall i = 1, ..., 15, \forall j = 1, ..., 15, \tag{3.7}$$

If $i = j$ in Equation (3.7) then it is called the direct arc elasticity, and otherwise the cross arc elasticity. In our application, the alternative scenario is a decrease of 20% of alternative j, $\Delta p_{jn} = -0.2 \cdot p_{jn}$. Results for each pair $(i, j), i = 1, \ldots, 15$, $j = 1, \ldots, 15$ are shown in Table 3.7.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-1.08	0.0163	0.0134	0.0133	0.0144	0.0121	0.0143	0.0145	0.0173	0.0133	0.0133	0.0138	0.0120	0.0139	0.0135
2	0.00836	-0.927	0.00553	0.00517	0.00725	0.00430	0.00672	0.00753	0.0151	0.00549	0.00524	0.00634	0.00427	0.00609	0.00566
3	0.0136	0.0111	-0.779	0.0182	0.0156	0.0197	0.0147	0.0147	0.00885	0.0176	0.0179	0.0163	0.0199	0.0160	0.0162
4	0.118	0.0902	0.161	-0.847	0.155	0.400	0.130	0.130	0.0648	0.161	0.165	0.148	0.186	0.144	0.146
5	0.0606	0.0623	0.0625	0.0719	-0.857	0.0701	0.0595	0.0606	0.0598	0.0626	0.0629	0.304	0.0637	0.0609	0.0603
6	0.0881	0.0620	0.138	0.291	0.123	-0.594	0.0995	0.103	0.0402	0.139	0.142	0.123	0.307	0.118	0.135
7	0.00758	0.00689	0.00773	0.00779	0.00766	0.00734	-1.35	0.00762	0.00647	0.00772	0.00778	0.00770	0.00730	0.00811	0.00766
8	0.000865	0.000881	0.000865	0.000871	0.000866	0.000847	0.000857	-0.722	0.000897	0.000865	0.000866	0.000866	0.000850	0.000861	0.000848
9	0.00464	0.00778	0.00230	0.00196	0.00347	0.00145	0.00332	0.00398	-1.40	0.00227	0.00211	0.00319	0.00158	0.00295	0.00237
10	0.00147	0.00118	0.00191	0.00199	0.00169	0.00216	0.00159	0.00160	0.000936	-0.707	0.00196	0.00178	0.00220	0.00174	0.00177
11	0.0250	0.0193	0.0335	0.0351	0.0293	0.0378	0.0275	0.0275	0.0148	0.0337	-0.822	0.0314	0.0478	0.0307	0.0309
12	0.00369	0.00341	0.00426	0.00438	0.0219	0.00455	0.00381	0.00386	0.00345	0.00427	0.00438	-1.00	0.00482	0.00408	0.00403
13	0.0532	0.0375	0.0846	0.0889	0.0685	0.187	0.0600	0.0625	0.0267	0.0856	0.104	0.0780	-0.339	0.0754	0.0802
14	0.0135	0.0117	0.0153	0.0157	0.0145	0.0160	0.0148	0.0141	0.0109	0.0154	0.0158	0.0150	0.0170	-0.951	0.0146
15	0.00845	0.00678	0.0103	0.0106	0.00943	0.0126	0.00904	0.00901	0.00527	0.0104	0.0106	0.00981	0.0122	0.00957	-2.34

Table 3.7: Direct and cross arc elasticities for each pair of alternatives.

[11] We also try multiple imputations, but the results do not change significantly, and considering it in this way saves time in the analysis.

The diagonal values are negative, as should be, and the off-diagonal values are positive. Therefore, decreasing the price of an alternative i increases the probability of choosing alternative i and decreases the probabilities of choosing all other alternatives. Moreover, by means of the cross nested logit, we get more realistic substitution patterns, compared to what we could obtain using a logit model. The ranges of the direct elasticities found are in line with what is reported by Train and Winston (2007b) (between -1.7 and -3.2 depending on the model they use). Berry et al. (1998) report direct elasticities that are a lot larger in absolute value, going up to -126 for some vehicles. However, the cross elasticities reported by Berry et al. (1998) are close to what we find. For example, the cross elasticity between the Mazda 323 (that belongs to the *medium* market segment) and the Nissan Maxima (that belongs to the market segment *full*) is reported to be 0.056. We obtain a value of 0.0136.

As an illustration of the substitution patterns obtained thanks to the CNL specification, we analyze the elasticities related to alternative 6, the *small diesel*. Figure 3.2 shows the values of the price arc elasticities obtained for alternative *small diesel*, namely $E^{\Delta P_n(6)}_{\Delta p_{jn}}$, $j = 1, \ldots, 15$. We see that when the price of *small diesel* is decreased, the largest arc cross elasticity is for alternative 13 (*small petrol*), followed by alternative 4 (*MPV diesel*). In other words, by making *small diesel* cars cheaper, we attract small petrol buyers more than any of the other vehicle car-types, followed by MPV-diesel. Due to the IIA property, this analysis could not be done with a logit model.

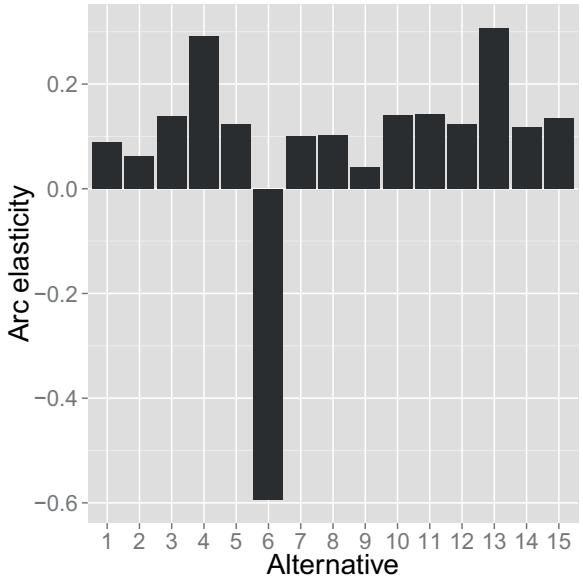


Figure 3.2: Price arc cross elasticities for medium diesel ($E^{\Delta P_n(6)}_{\Delta p_{jn}}$).

3.6.2 Comparing different future scenarios

For the forecasting exercise we consider three scenarios. The first one, denoted by *do nothing* scenario, corresponds to a foreseeable future where no specific policy is implemented. The second one, denoted the *tax* scenario, uses the same assumptions as the *do nothing* scenario, plus an increase of the registration tax for internal combustion vehicles of 10% and an increase in the fuel price. Finally, the *technological innovation* scenario uses the same assumptions as the *do nothing* scenario, plus a decrease in the price of electric vehicles of 15% and an increase of the range of these vehicles by 100%. They are all considered to be related to a five-years horizon. The socioeconomic characteristics of new car buyers are assumed to remain unchanged.

The mean value of fuel consumption decreased from $6.95l/100km$ in 2010 to $6.49l/100km$ in 2015 (Comité des Constructeurs Français d'Automobiles, 2016). This represents a 7% decrease. We assume that the decrease in a five-years time horizon will be the same. For the price, in the *do nothing* scenario, motivated by the decrease of the bonus-malus in France from 2015 to 2016 from 4000€ to 1000€ for rechargeable hybrids and from 2000€ to 750€ for other hybrids (Ministère de l'environnement, de l'énergie et de la mer, 2016), we assume an increase of 2500€ of the price of all hybrid vehicles. Moreover, for the *tax* scenario we assume that an increase in the registration tax will render internal combustion vehicles 10% more expensive. For the *technological innovation* scenario we assume that an improvement in the manufacturing process will render electric vehicles 15% cheaper. For the electric vehicles' range in both the *do nothing* and the *tax* scenarios, we assume that within five years the range will increase of 50km for all vehicles. This is in line with the ranges for the Nissan leaf comparison between 2011 (117km) and 2016 (172km) (U.S. Department of Energy, 2016). For the *technological innovation* scenario we assume that the ranges for all electric vehicles on the market are doubled. Finally, for the fuel price, we assume that the petrol and diesel prices will be the same, as the French government has reported that they would like the difference between both prices to decrease (Sud Ouest, 2015). We assume that the taxes are constant in the *do nothing* and *technological innovation* scenarios, and use the forecast for the price of imported fuel (European Comission, 2011), resulting in $2.44€/l$. For the *tax* scenario we use the same import price, and increase the taxes by 50% which results in $3€/l$. These assumptions are summarized in Table 3.8.

We compute the market shares of each alternative for each scenario. They are reported in Table 3.9. In order to interpret these results, we focus on the electric vehicle alternative and plot the market shares per income level and scenario. This is shown in Figure 3.3. There are 11 income levels as shown in Table 3.3 and they are labeled from *Income 1* for the lowest income level, to *Income 11* for the highest. Indeed, all the scenarios have an increase in the share of new sold electric vehicles, and the most effective scenario is the one with a major technological advance. It is also very interesting to note how

	Do nothing scenario	**Tax scenario**	**Technological innovation scenario**
Max. power	-	-	-
Fuel cons.	7% decrease in fuel consumption (alt 1-14)	7% decrease in fuel consumption (alt- 1-14)	7% decrease in fuel consumption (alt. 1-14)
Price	- Hybrid vehicles (alt 7 and 14) 2500 € more expensive	- Hybrid vehicles (alt. 7 and 14) 2500 € more expensive - Internal combustion engine vehicles (all alt. except 7,14,15) 10% more expensive	- Hybrid vehicles (alt. 7 and 14) 2500 € more expensive - Electric vehicles (alt. 15) 15% cheaper
Range	+50km	+50km	100% increase
Fuel price	diesel=petrol=2.44€/ℓ	diesel=petrol=3€/ℓ	diesel=petrol=2.44€/ℓ

Table 3.8: Description of the different tested scenarios.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Do nothing	1.15	0.521	2.32	23.2	7.27	30.0	0.511	0.115	0.134	0.288	4.45	0.570	27.3	1.57	0.648
Tax	1.08	0.492	2.27	22.6	6.96	30.6	0.531	0.109	0.120	0.281	4.27	0.546	27.8	1.65	0.725
Techno. innov.	1.15	0.525	2.32	23.2	7.38	29.0	0.508	0.121	0.139	0.293	4.56	0.593	27.6	1.55	1.09
Base	1.16	0.528	2.35	23.5	7.46	29.6	0.554	0.119	0.137	0.289	4.49	0.579	27.1	1.65	0.486

Table 3.9: Predicted market shares for each alternative and scenario in percentages.

the increase in market share is higher for medium income levels rather than low or high income levels. In other words, people with lower levels of income can still not afford the electric vehicles, while people with higher levels of income could also afford them before the decrease in price, so are less attracted by this improvement. In most studies in the literature, only the fixed effect of income is considered (income enters linearly the utility function). Since the interaction between income and price is not included, this behavior cannot be captured by them.

We repeat the analysis for hybrid vehicles. Figure 3.4(a) shows the hybrid diesel and Figure 3.4(b) shows the hybrid petrol. In both cases, the share is a growing function of the income. However, for the *hybrid diesel* alternative, the market shares in the defined scenarios actually decrease for all income levels. This indicates that without the subsidies given today, the sales of hybrid diesel cars would decrease even if internal combustion engines become 10% more expensive. For the petrol case, with the *tax* scenario, the market shares increase slightly – or stay the same– for all income levels, but they decrease in both the *do nothing* and the *technological innovation* scenarios. We can conclude that unless internal combustion engine (ICE) cars are made more expensive (like in the *tax* scenario), the combination of increasing the fuel price and decreasing the subsidies for hybrid vehicles does not allow the new sales market shares of these alternatives to increase. Wang et al. (2016) show that the consumers' attitudes towards hybrid electric vehicles influence the adoption intention of these types of vehicles. This could be introduced in our framework by introducing an Integrated Choice and Latent Variable model (ICLV), but is considered future work.

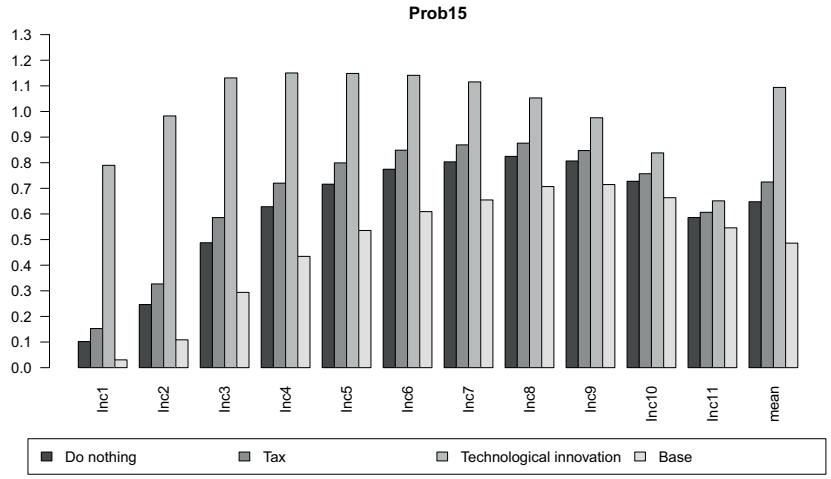


Figure 3.3: Market shares for the electric vehicle alternative for the base case and each of the scenarios, per income level.

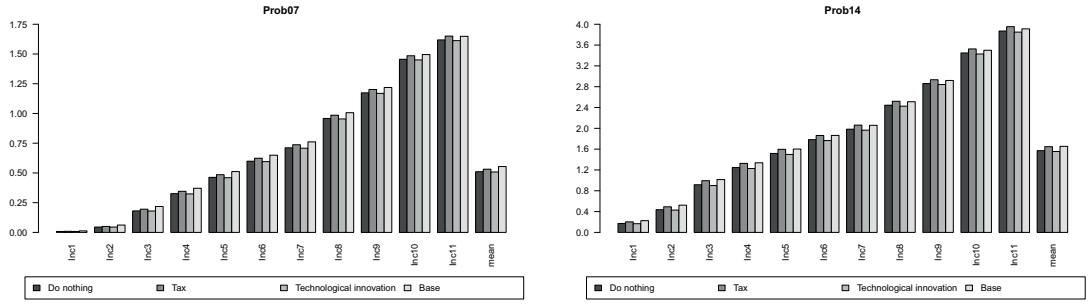
It is important to note that these predicted market shares can only be calculated using revealed preference data for the sample enumeration. Indeed, the values of the attributes in stated preferences data are engineered by the experimental design, and do not represent any market reality. Note also that the shares refer to the reference population, that is the set of people buying a new car during a given year.

We do not model the duration of car ownership. To do so, we would need to model the dynamics (de Lapparent and Cernicchiaro, 2012; Cernicchiaro and de Lapparent, 2015), and we cannot, because we have cross sectional data. Adda and Cooper (2000) and de Palma and Kilani (2008) show that some policies can have counterintuitive effects when modeling car replacement. Subsidizing the renewal of old cars slows down the renewal of cars, so the impact of the policy is not necessarily positive: pollution is higher and the market of new cars is harmed. In the same sense, increasing the taxes of gasoline leads to lower mileages, but larger replacement times.

3.6.3 Willingness to pay

We compute the marginal willingness to pay for an increase of maximum power and the willingness to accept an increase in fuel consumption, which have the following expressions

$$\begin{aligned} \text{WTP}(max\_power_{in}) &= -\frac{\frac{\partial V_{in}}{\partial max\_power_{in}}}{\frac{\partial V_{in}}{\partial price_{in}}} = -\frac{\beta_{max\_power} \cdot income}{100 \cdot \beta_{price\_inc\_i}} \quad \left[\frac{€}{bhp}\right], \\ \text{WTA}(cons_{in}) &= \frac{\frac{\partial V_{in}}{\partial cons_{in}}}{\frac{\partial V_{in}}{\partial price_{in}}} = \frac{10 \cdot price(fuel) \cdot \beta_{conso\_inc}}{\beta_{price\_inc\_i}} \quad \left[\frac{€}{\ell/km}\right]. \end{aligned} \tag{3.8}$$



(a) Hybrid diesel

(b) Hybrid petrol

Figure 3.4: Market shares for hybrid vehicles for the base case and each of the scenarios, per income level.

Results are summarized per alternative in Table 3.10. By considering an individual who drives 13000km (which is the mean mileage in France in 2014 for private vehicles (Institut national de la statistique et des études économiques, 2016a)) and who keeps the car for 5.4 years (Institut national de la statistique et des études économiques, 2016b), the real savings would be of 906€ for the diesel cars and of 1040€ for petrol cars. The results of the model show comparable willingness to pay values, ranging from 587€, to 1630€, for a decrease of $1\ell/100km$ in the fuel consumption. For the maximum power our results show that an individual is willing to pay 233€ more for a car that has 1bhp more of maximum power, if all else is equal. Paying hundreds of euros for an additional bhp of power is only observed in the market for extreme versions of cars (e.g.: A3 Sportback 118bhp costs 26260€, and RS3 Sportack 367bhp costs 59860€, which makes 135€/bhp (Autobild, 2017)). Our model provides higher willingness to pay for engine maximum power that what is usually observed in the market. It is future work to investigate this further. A possible direction is to include alternative-specific coefficients in the utilities for the engine maximum power.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Mean
Fuel cons. [€/ (ℓ/100km)]	817	1020	826	1090	1000	1200	587	1630	921	874	1210	1160	1590	865	-	1060
Max power [€/ bhp]	204	255	207	274	251	300	147	356	201	191	264	252	347	189	542	233

Table 3.10: Willingness to accept for fuel consumption and willingness to pay for maximum power for each alternative.

3.7 Conclusions and future work

We have developed a CNL model for car type choice for new buyers, using revealed preference data. A multiple imputations method has been applied for the attributes of

the non chosen alternatives. We have used the estimated model to study the effects of several policy scenarios in the market shares of different car-types with a special focus on electric vehicles. The results are in line with the market share variations obtained in other models using revealed preference data (but not focusing on electric vehicles). We also computed price elasticities, that are in line with values found in the literature, and willingness to pay and to accept values, that are in line with what is observed in the new vehicle market.

By using revealed preference data we have encountered two major difficulties. The first relates to the definition of the choice set. We aggregated several make-model-type of cars into fifteen alternatives defined as a combination of a market segment (full, luxury, medium, MPV, offroad and small) and a fuel type (diesel, petrol, hybrid, electric). This definition of the alternatives makes it natural to estimate a cross nested logit model, since alternatives that share either market segment or fuel type share unobserved attributes. The results confirm that the cross nested logit model is better than the logit model both in terms of fit and in terms of realistic behavioral results.

The second major difficulty of using RP data is related to the definition of the attributes of the unchosen alternatives. To the best of our knowledge, it is the first time that the empirical distributions of the attributes of the alternatives are used to impute the attributes of the unchosen alternatives. This is computationally slow, but much more realistic than considering the mean of the attributes of the chosen alternatives, as was done in the literature in the past.

This methodology, however, is not free of limitations. As for any choice experiment, we are not able to estimate a parameter if the attribute related to it has very little variability. In our data, this is the case for the range of electric vehicles. EVs represent a very small part of the 2014 car market in France, and therefore, the reported ranges do not allow to estimate the sensitivity to the autonomy of electric vehicles. In order to overcome this problem, we use the willingness to pay for range reported in the literature.

Out of the three tested scenarios, the most effective in order to increase the use of electric vehicles is the *technological innovation* scenario. The results we obtain are also divided per income levels. The middle-income levels are the ones that would increase the market share of EV the most.

Some improvements in the presented research would include weighting the alternatives by the amount of different number of vehicles that they contain, as done in Train (1986). Moreover, the variables related to the attributes of the alternatives could contain measurement errors, since they consist of reported values (instead of catalog values), which might cause endogeneity. Auxiliary models for the car attributes can be estimated and integrated with the choice model to solve this issue. Our framework allows for these auxiliary models to be included easily. Moreover, these auxiliary models would also

allow to recover observations containing missing attributes. Another future research direction would be to take into account the price endogeneity as in Berry et al. (1995). Furthermore, new results by Mai et al. (2015) show that it might be feasible to estimate the model over the full set of alternatives, without the need for either aggregation or sampling of alternatives. It would also be interesting to estimate the same model over the full set of alternatives and compare the results obtained with the two approaches. In the same direction, it would also be interesting to compare our method to impute the attributes of the aggregate alternatives to the methods examined and compared in Wong et al. (2017), in order to identify any possible biases in the parameter estimation.

4

Discrete-continuous maximum likelihood

This chapter is based on the technical report:

Fernández-Antolín, A., Lurkin, V., Bierlaire, M. (2017). Discrete-continuous maximum likelihood for the estimation of nested logit models. Working paper. The work has been performed by the candidate under the supervision of Dr. Virginie Lurkin and prof. Michel Bierlaire.

4.1 Introduction

In the estimation of discrete choice models, in general, only continuous parameters are considered, although some models include also discrete ones. The most typical example is the nest membership parameter of a nested logit model. In some cases, it is clear which alternatives share unobserved attributes and the nesting structure is obvious. However, in other cases there are several nesting structures that make intuitive sense. In practice, to determine the most appropriate nesting structure, the analyst has several options: (i) to enumerate all the possible values, and estimate the continuous parameters for each combination, and (ii) to make the problem continuous by relaxing the integrality of the discrete parameters. For instance, a membership indicator becomes a continuous parameter between 0 and 1 (like in the cross nested logit model), or by making the membership probabilistic (like in latent class models). In both cases, however, the

likelihood function features several local optima, so that classical nonlinear optimization methods may not find the (global) maximum likelihood estimates.

In this chapter, we propose a new mathematical model that is designed to find the global maximum likelihood estimates of a choice model involving both discrete and continuous parameters. We call our approach *discrete-continuous maximum likelihood* (DCML) because we introduce into the maximum likelihood framework binary parameters. We build on on the framework developed by Pacheco et al. (2017) to formulate our problem as a mixed integer linear problem (MILP), because they can be solved to (global) optimality. This is a first attempt towards a complete MILP formulation of the maximum log-likelihood, which results in a problem with high computational complexity. The goal of this chapter is to show under which circumstances our approach is computationally feasible, and to study its potentials and limitations. To do so, we use an example of stated preference data with three alternatives

Our contributions are multiple. First, to the best of our knowledge, this is the first time that discrete parameters are estimated and included in the maximum likelihood framework in the context of discrete choice models. Second, our model is formulated as an MILP. We use simulations and piecewise linear function approximation to dispose of the non-linearity of the log-likelihood. To the best of our knowledge, this is the first time that the log-likelihood is linearized. Finally, the proposed mathematical model is general and can be used with any choice model, as long as the distribution of the error terms can be simulated (e.g.: cross nested, logit or latent class models).

The remaining of the chapter is organized as follows. The literature review is presented in Section 4.2, followed by the mathematical model in Section 4.3. The case study is presented in Section 4.4 and the conclusions of the paper and future research directions are presented in Section 4.5.

4.2 Literature review

The most typical example of a discrete parameter that is usually disregarded from the estimation process is the nest allocation parameter in nested logit models. Nesting structures are used in discrete choice models when correlation between alternatives is suspected. They are used in a very broad range of contexts such as airline itinerary choice (Lurkin, 2016), car-type choice (as discussed in section 3.2), route choice (Vovsha and Bekhor, 1998), residential location choice (Zolfaghari, 2013), and in mode choice (Koppelman and Bhat, 2006; Forinash and Koppelman, 1993) among others. In route choice, for instance, two paths are correlated if they share a physical segment of the route. However, in other contexts, the partition of alternatives into different nests is less obvious and researchers either decide *a priori* which is the optimal nesting structure,

or enumerate some of them and choose the best one *a posteriori*. Since the number of nesting possibilities increases combinatorially with the number of alternatives, it is often not feasible to enumerate all of them.

In the context of parking-location choice, for example, Chaniotakis and Pel (2015) and Hunt and Teply (1993) predefine the nesting structures, and suggest that the nests are defined by *on street* and *off street* parking. They don't study other possible nesting structures. In the context of access mode and airport choice, Pels et al. (2003) consider several nesting structures, based on common airport or common access mode, and decide *a posteriori* the most adequate one for their case study. In a flight-route choice model, Yang and Wang (2017) argue that similarities between alternatives (and therefore nesting structures) could derive from sources like origin airport, destination airport, market share, airport capacity, and/or access distance. They estimate several of them and report only the best one. Coldren and Koppelman (2005) also use the *a posteriori* approach for a air-travel itinerary choice case study, as does Lurkin (2016).

In car-type choice, Hoen and Koetse (2014) try several nesting structures based on fuel type, and conclude that they are not better than the logit model, while Berkovec and Rust (1985) predefine the nesting structure based on vehicle size and age categories. They justify their nesting decision as being aligned with the automobile industry market classification, and do not test other nesting structures. McCarthy and Tay (1998) also predetermine their nesting structure, but nests are defined by fuel efficiency levels.

In discrete choice models, the logit has a concave log-likelihood function (as long as the model is linear-in-parameters) and therefore a global optimum exists. However this is not the case for nested logit (Daganzo and Kusnic, 1993), cross nested logit (Bierlaire, 2006; Abbe et al., 2007) or latent class models. Knockaert (2015) discusses the problem of local optima in latent class models. Ordered generalized extreme value models (OGEV) also have non-concave log-likelihood functions. Lurkin (2016) identifies in her thesis that some of the estimated OGEV models converge to local optima. So far, the way to tackle this problem in the literature has been to develop heuristics where several starting points of the optimization algorithm are tested (Bierlaire et al., 2009; Boeri, 2011; Hole and Yoo, 2017). By linearizing the maximum likelihood function, our problem is formulated as a MILP for which exact algorithms exist that can solve the problem to optimality. Since the framework relies on the simulation of the error terms, this is true for any model for which we are able to simulate the error terms (also when there are no discrete parameters involved).

4.3 Mathematical model

In this section, we derive the maximum likelihood problem as an MILP, for which global optimality can be reached, and use it to estimate simultaneously discrete and continuous parameters[12]. It is organized as follows. In Section 4.3.1 we develop the linearization of the maximum likelihood function. Then, in Section 4.3.2, we use the nested logit model as an illustration of how to introduce the estimation of discrete parameters in this framework. In this context, we estimate the binary allocation parameters of alternatives to nests. Finally, in Section 4.3.3 we discuss the complexity of the presented mathematical problem.

4.3.1 Linearization of the maximum likelihood

Let's define the utility function of an individual n for an alternative i as follows

$$U_{in} = V_{in} + \varepsilon_{in}, \qquad \forall n = 1, \ldots, N\,, \forall i = 1, \ldots, J, \tag{4.1}$$

where

- N is the number of individuals in the sample,
- J is the number of alternatives,
- $V_{in} = \sum_k \beta_k x_{in}^k$ is the deterministic part of the utility for alternative i and individual n, where
  - β_k are the parameters to be estimated,
  - x_{in}^k is the value of the kth variable for alternative i and individual n.
- ε_{in} are the error terms.

The choice model resulting from this specification depends on the distributional assumption of ε_{in}. In order to estimate the parameters of the model, we apply maximum likelihood. That is, we maximize the following function

$$\log\left(\prod_{n=1}^{N}\prod_{i\in\mathcal{C}_n} P_n(i)^{d_{in}}\right), \tag{4.2}$$

where

[12]Note that in the field of discrete choice models, the term *parameter* is used for what is estimated, while the term *variable* is used to define observed data. In the field of operations research it is the opposite: the term *variable* or *decision variable* are the outcomes of the optimization problem, while the *parameters* are observed. In this chapter, we follow the terminology from discrete choice models.

- $P_n(i)$ is the probability that individual n chooses alternative i,
- d_{in} is observed and takes value 1 if individual n chooses alternative i and 0 otherwise,
- $\mathcal{C}_n$ denotes the choice set of individual n, and is defined as

$$\mathcal{C}_n = \{i \,|\, y_{in} = 1, i = 1, \ldots, J\}, \qquad \forall n, \tag{4.3}$$

  where y_{in} represents the availability of an alternative i for an individual n, and is observed from the data. y_{in} takes value 1 if alternative i is available to individual n and 0 otherwise.

As mentioned above, the aim is to convert this problem in an MILP. We therefore need to address the two sources of nonlinearities from Equation (4.2), namely, (i) the expression of the probabilities (that are highly non-linear and non-concave), and (ii) the logarithm function.

Linearization of the expression of the probabilities In order to linearize the expressions of the probabilities, we rely on the framework developed by Pacheco et al. (2017), and we generate R draws based on the distributional assumption of ε_{in} from Equation (4.1), $\varepsilon_{in1}, \ldots, \varepsilon_{inR}$. The realization of a draw r is referred to as *scenario* r. Then, the choice of an individual n in a particular scenario r is characterized by the following binary parameters

$$w_{inr} = \begin{cases} 1 & \text{if} \quad U_{inr} > U_{jnr} \quad \forall j \neq i, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall i \in \mathcal{C}_n, n, r. \tag{4.4}$$

U_{inr} denotes the utility function for alternative i and individual n in scenario r and has the following expression

$$U_{inr} = V_{in} + \varepsilon_{inr}, \qquad \forall n = 1, \ldots, N\,, \forall i \in \mathcal{C}_n\,, \forall r = 1, \ldots, R. \tag{4.5}$$

The linearization of Equation (4.4) together with additional constraints to ensure that the choices are well defined have been developed by Pacheco et al. (2017) and are described in Appendix A. Note that the objective of the model by Pacheco et al. (2017) and ours is different. Their objective is to have a unified model of demand and supply. They use a discrete choice model as the demand model, and this is their motivation to linearize the expressions of the probabilities. Our approaches differ in that their objective function is benefit (or revenue) maximization, and their decision variables are the price levels. They consider the parameters in the discrete choice model as given, and fix them based on values reported in the literature. We investigate how to use part of

their framework to convert the problem of estimation, through maximum likelihood, in an MILP.

This formulation allows to express the probability of individual n to choose alternative i as the average of w_{inr} over scenarios, as follows

$$\hat{P}_n(i) = \frac{1}{R}\sum_{r=1}^{R} w_{inr}, \qquad \forall i \in \mathcal{C}_n, n. \tag{4.6}$$

Then, by substituting $P_n(i)$ by its approximation from Equation (4.6) in Equation (4.2) the objective function becomes

$$\sum_{n=1}^{N}\sum_{i\in\mathcal{C}_n} d_{in}\left(\log\left(\sum_{r=1}^{R} w_{inr}\right) - \log(R)\right). \tag{4.7}$$

Linearization of the logarithm The only remaining non-linearity is now the logarithm that appears in the objective function. Notice that $\sum_{r=1}^{R} w_{inr}$ can only take integer values from 0 to R, depending on the number of draws for which the utility associated with individual n and alternative i is the highest (i.e.: in how many scenarios individual n chooses alternative i). We define

$$s_{in} = \sum_{r=1}^{R} w_{inr}, \tag{4.8}$$

and

$$z_{in} = \log(s_{in}). \tag{4.9}$$

We can approximate the logarithm function with a piecewise linear function so that both are equal at the integer values. Since at the integer values both functions are equal, and that we need to evaluate the logarithm only at integer values, the specification with the logarithm and the specification with the piecewise approximation are equivalent. In order to define this piecewise linear function, we denote by $PL^p(s_{in})$ the line that passes through points $(p-1, \log(p-1))$ and $(p, \log(p))$, $\forall p = 1, \ldots, R$[13], then $PL^p(s_{in})$ has the following expression:

$$PL^p(s_{in}) = \log(p)(s_{in} - (p-1)) + \log(p-1)(p - s_{in}), \quad \forall p = 1, \ldots, R. \tag{4.10}$$

[13]Since $\log(0)$ is not defined, and $\lim_{x\to 0}\log(x) = -\infty$ we approximate $-\infty$ by a *negative enough* number and denote it L_0. In practice we consider $L_0 = -100$.

Therefore, the maximization of Equation (4.7) is equivalent to

$$\max \quad \sum_{n=1}^{N}\sum_{i=1}^{J} d_{in}\left(\sum_{r=1}^{R} z_{in} - \log(R)\right) \tag{4.11}$$

$$\text{subject to} \quad s_{in} = \sum_{r=1}^{R} w_{inr} \qquad \forall i, n \tag{4.12}$$

$$z_{in} \leq PL^{p}(s_{in}) \qquad \forall i, n, p \tag{4.13}$$

Note that the objective functions (4.7) and (4.11) are not equivalent, but the optimal solutions of both problems are the same due to the fact that it is a maximization problem. Figure 4.1 shows the relation between s_{in}, z_{in}, and $\log(s_{in})$ in a schematic way.

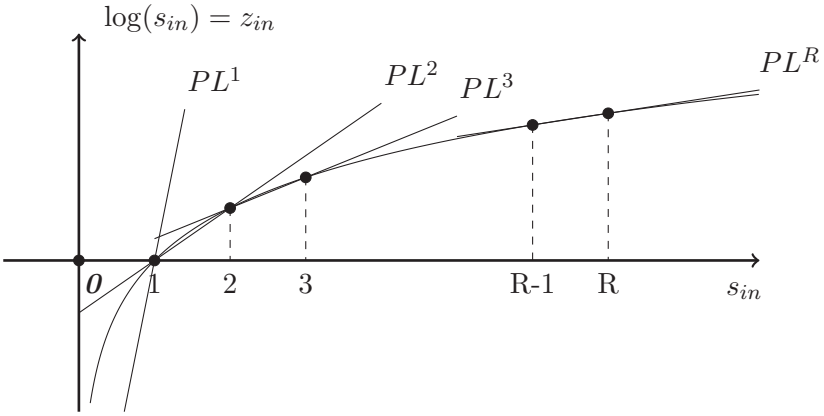


Figure 4.1: Relation between s_{in} and z_{in}

4.3.2 Adapting the nested logit

The mathematical model presented in this section is a particular example of the framework presented above for the nested logit model. We show how discrete parameters can be estimated simultaneously with the continuous ones. The framework remains valid for any DCM for which draws can be generated.

Following Ben-Akiva and Lerman (1985), the error term associated with each alternative i that belongs to nest m can be decomposed into a common error component, ε_{mn}, and an independent error term, ε_{imn}

$$\varepsilon_{in} = \varepsilon_{mn} + \varepsilon_{imn}, \tag{4.14}$$

where

- ε_{mn} is such that $\tilde{\varepsilon}_{mn} = \varepsilon_{mn} + \varepsilon'_{mn}$, and

  - $\tilde{\varepsilon}_{mn} \overset{iid}{\sim} EV(0, \mu)$,

  - $\varepsilon'_{mn} \overset{iid}{\sim} EV(0, \mu_m)$.

- $\varepsilon_{imn} \overset{iid}{\sim} EV(0, \mu_m)$,

Therefore, Equation (4.1) can be rewritten as

$$U_{in} = V_{in} + \tilde{\varepsilon}_{mn} + (\varepsilon_{imn} - \varepsilon'_{mn}). \tag{4.15}$$

For normalization reasons, we assume that $\mu = 1$ and that therefore $\mu_m \geq \mu = 1$.

From the properties of the extreme value distribution, we know that if $\varepsilon_{imn} \sim EV(0, \mu_m)$, then

$$\varepsilon_{imn} = \frac{1}{\mu_m} \xi_{imn}, \tag{4.16}$$

verifies that $\xi_{imn} \overset{iid}{\sim} EV(0, 1)$. Analogously, if $\varepsilon'_{mn} \sim EV(0, \mu_m)$, then

$$\varepsilon'_{mn} = \frac{1}{\mu_m} \xi'_{mn}, \tag{4.17}$$

verifies that $\xi'_{mn} \overset{iid}{\sim} EV(0, 1)$. Equation (4.15) can therefore be rewritten as

$$U_{in} = V_{in} + \tilde{\varepsilon}_{mn} + \frac{1}{\mu_m} (\xi_{imn} - \xi'_{mn}). \tag{4.18}$$

Finally, as we don't know a priori if alternative i belongs to nest m, we introduce the following indicator parameters:

$$b_{im} = \begin{cases} 1 & \text{if alternative } i \text{ belongs to nest } m, \\ 0 & \text{otherwise,} \end{cases} \tag{4.19}$$

that sum up to one (to express that each alternative belongs to exactly one nest). This is imposed with the following constraint

$$\sum_{m=1}^{M} b_{im} = 1, \qquad \forall i. \tag{4.20}$$

Then, the utility (4.18) becomes:

$$U_{in} = V_{in} + \sum_{m=1}^{M} b_{im} \left(\tilde{\varepsilon}_{mn} + \frac{1}{\mu_m}(\xi_{imn} - \xi'_{mn}) \right) \tag{4.21}$$

$$= V_{in} + \sum_{m=1}^{M} b_{im}\tilde{\varepsilon}_{mn} + \sum_{m=1}^{M} \left(\frac{b_{im}}{\mu_m}(\xi_{imn} - \xi'_{mn}) \right). \tag{4.22}$$

In order to linearize Equation (4.22), we define

$$\bar{\mu}_m = \frac{1}{\mu_m} \in (0, 1], \tag{4.23}$$

and

$$\tau_{im} = b_{im}\bar{\mu}_m. \tag{4.24}$$

Then, the linearization of Equation (4.22) is classic, and is as follows

$$U_{in} = \sum_{k} \beta_k x_{in}^k + \sum_{m=1}^{M} b_{im}\tilde{\varepsilon}_{mn} + \sum_{m=1}^{M} \tau_{im}(\xi_{imn} - \xi'_{mn}), \qquad \forall i \in \mathcal{C}_n, n, \tag{4.25}$$

$$\tau_{im} \leq b_{im}, \qquad \forall i, m, \tag{4.26}$$

$$\tau_{im} \leq \bar{\mu}_m, \qquad \forall i, m, \tag{4.27}$$

$$\tau_{im} \geq \bar{\mu}_m + b_{im} - 1, \qquad \forall i, m, \tag{4.28}$$

$$\tau_{im} \geq 0, \qquad \forall i, m. \tag{4.29}$$

For identification purposes, the scale of a nest with only one alternative must be 1. Moreover, if a nest m contains no alternatives, then the value of $\bar{\mu}_m$ can be set to any value (since it does not affect the objective function). We arbitrarily decide to fix $\bar{\mu}_m$ to 1 if nest m is empty. That is,

$$\text{if} \quad \sum_{i=1}^{J} b_{im} \leq 1 \quad \text{then} \quad \bar{\mu}_m = 1, \qquad \forall m. \tag{4.30}$$

To linearize this implication, we introduce binary parameters q_m that take value 1 if $\sum_{i=1}^{J} b_{im} \leq 1$. The linearization is then as follows

$$\sum_{i=1}^{J} b_{im} \geq 2(1 - q_m), \qquad \forall m, \tag{4.31}$$

$$\bar{\mu}_m \geq q_m, \qquad \forall m. \tag{4.32}$$

To prove the equivalence we consider the following:

- If $\sum_{i=1}^{J} b_{im} = 0$, constraints (4.31) become

$$0 \geq 2(1 - q_m), \qquad \forall m, \tag{4.33}$$

  which is equivalent to $q_m \geq 1$, and since it is a binary parameter, we obtain that $q_m = 1$. Then, from Equation (4.32) we have that $\bar{\mu}_m \geq 1$. Since by definition, $\bar{\mu}_m \in (0, 1]$, we obtain that $\bar{\mu}_m = 1$.

- If $\sum_{i=1}^{J} b_{im} = 1$, constraints (4.31) become

$$1 \geq 2(1 - q_m), \qquad \forall m, \tag{4.34}$$

  which is equivalent to $q_m \geq 0.5$, and since it is a binary parameter, we obtain that $q_m = 1$. Analogously, we obtain that $\bar{\mu}_m = 1$.

- If $\sum_{i=1}^{J} b_{im} \geq 2$, Equation (4.31) is always true (both with $q_m = 0$ and with $q_m = 1$), therefore q_m is free, and Equation (4.32) is always verified.

In a nested logit model, a maximum of $J-1$ alternatives can belong to a nest. Therefore we also need to add the constraint that

$$\sum_{i=1}^{J} b_{im} \leq J - 1, \qquad \forall m. \tag{4.35}$$

The definition of b_{im} from Equation (4.19) leads to several combinations of b_{im} resulting in the same nesting structure. Consider an example with three alternatives and three nests. Table 4.1 shows the possible combinations of b_{im} where alternatives 1 and 2 are correlated, and alternative 3 is not.

i \ m	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
2	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
3	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	0

Table 4.1: Values of b_{im} that render equivalent nesting structures.

In order to reduce the possible number of combinations that result in the same nesting structure, we impose that

$$b_{im} = 0, \qquad \forall m > i. \tag{4.36}$$

Then, only the first two configurations of the b_{im} parameters are possible, therefore reducing the solution space.

Note that different distributional assumptions in Equation (4.5) lead to different choice models, such as logit or error component models. For example, we can use this framework to estimate the parameters of a logit model by considering that $\varepsilon_{inr}, r = 1, \ldots, R$, are R draws of an $EV(0, 1)$ distribution. The details of how to adapt the framework to error component models is described in Appendix B.

4.3.3 Complexity of the problem

The model presented above has a large number of constraints and variables to be estimated, that increase as a function of the number of draws R, the number of respondents N, and the number of alternatives J. To solve the problem we use the CPLEX Interactive Optimizer (CPLEX version 12.7.0.0), which is a standard solver. However, specialized algorithms relying on decomposition methods are necessary when the problem become large. This is out of the scope of this chapter, but needs to be investigated further.

Moreover, our formulation contains *big M* constraints (see Constraint (A.5) from Appendix A), which depend on the upper and lower bounds of the utility functions. The tighter the bounds on β_k, the tighter the formulation, and the faster it is to solve the problem. Due to this constraint, the alternative specific constants of the nested logit models must be set to zero, as otherwise the problem cannot be solved.

Previous formulations of the model, that proved to be slower to solve, can be found in Appendix C.

4.4 Case study

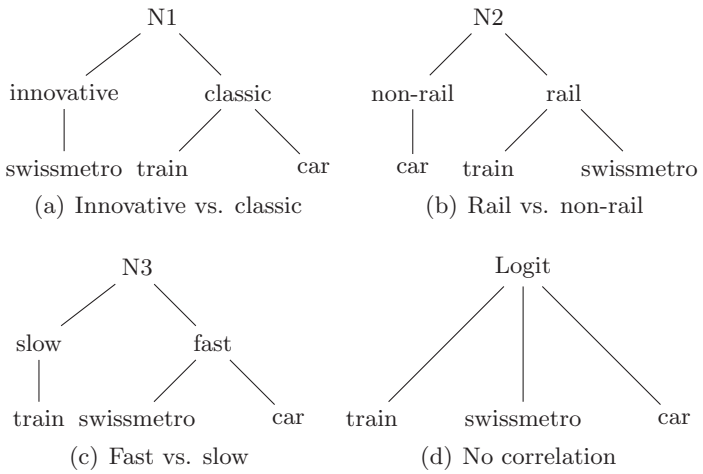
In this case study we first examine the simulation error in the value of the final log-likelihood for given values of the parameters. Then, we investigate the strengths and limitations of the MILP presented above by using it to estimate (i) the continuous parameters of a logit model, and (ii) the continuous and discrete parameters of a nested logit model. To do so, we use a stated preferences mode choice case study collected in Switzerland in 1998. The respondents provided information in order to analyze the impact of the modal innovation in transportation represented by the Swissmetro, a maglev underground system, compared to the usual transport modes of car and train.

The choice set of the respondents is $\mathcal{C} = \{\text{car, train, swissmetro}\}$. Using logit and nested logit models there are four possible nesting structures, involving different assumptions:

- The modes *train* and *car* share unobserved attributes due to the fact that they

are both *classic* or *existing* transportation modes. The corresponding nesting structure is represented in Figure 4.2(a).

- The modes *train* and *swissmetro* share unobserved attributes, due to the fact that they are both rail-based, unlike the *car* alternative. The corresponding nesting structure is represented in Figure 4.2(b).
- The modes *swissmetro* and *car* are correlated due to the fact that they are generally perceived as faster than the *train* alternative. The corresponding nesting structure is represented in Figure 4.2(c).
- There is no correlation between the different alternatives. The corresponding nesting structure is represented in Figure 4.2(d).



(a) Innovative vs. classic (b) Rail vs. non-rail

(c) Fast vs. slow (d) No correlation

Figure 4.2: Possible nesting structures with two nests.

The number of respondents in the dataset is 1192, with 9 response tasks each. However, in order to apply our framework, we consider a subset of 200 observations, which is the empirical minimum so that the nesting structures are identified.

Table 4.2 shows the model specification considered, and Table 4.3 shows the estimation results of each of the nesting structures, with this model specification. Note that the model corresponding to N3 is not identified. Table 4.4 shows the estimation results for the nested nested logit model when the alternative specific constants are fixed to zero. Note that models N2 and N3 are not identified. These estimation results are calculated using a state-of-the-art continuous estimation software: PyhtonBiogeme (Bierlaire, 2016).

As discussed in Section 4.3.3, the tightness of the problem depends on the value of M_{nr} (see Constraint (A.5) from Appendix A), that depends on the upper and lower bounds given to the parameters to be estimated. They are summarized in Table 4.5. We use

Parameter	Car	Train	Swissmetro
ASC_{CAR}	1	0	0
ASC_{SM}	0	0	1
β_{COST}	cost	cost	cost
β_{TIME}	time	time	time

Table 4.2: Model specification - Deterministic part of the utility functions

	N1		N2		N3		Logit	
	param.	p-value	param.	p-value	param.	p-value	param.	p-value
ASC_{CAR}	-0.0287	0.63	-1.32	0.00	-0.545	0.04	-0.545	0.04
ASC_{SM}	0.574	0.00	0.182	0.43	0.778	0.00	0.778	0.00
β_{COST}	-0.0581	0.63	-0.261	0.26	-0.376	0.10	-0.376	0.10
β_{TIME}	-0.0853	0.62	-0.309	0.06	-0.364	0.10	-0.364	0.10
$\bar{\mu}$	0.0403	0.62	0.362	0.17	1.00	1.00	-	-
FLL	-155.8		-161.4		-162.2		-162.2	

Table 4.3: Estimation results using the state-of-the-art continuous estimation of the different nesting structures.

ranges of 0.5[14] for all the parameters such that they include the values obtained by the continuous estimation software. A limitation of the MILP is that for the nested logit model, the values of ASC_{CAR} and ASC_{SM} are fixed to zero since otherwise the problem cannot be solved in reasonable time.

4.4.1 Investigating the simulation error

As the proposed framework relies on simulation, it is important to start by determining the minimum number of draws needed to obtain reliable values of the final log-likelihood. To do so, we evaluate Equation (4.11) at the values of the parameters obtained by the continuous estimation software that are reported in Table 4.3. We do so for the logit model, and for the three possible nested logit models (N1, N2 and N3). The results are shown in Table 4.6, together with the value of the final log-likelihood (FLL) obtained with the continuous estimation. The table also shows the relative error between the real FLL and the value obtained with the MILP. Let FLL be the true value of the final log-likelihood, and $\widehat{FLL}_R$ the value obtained for R draws. The relative error e^R_{FLL} is

[14]Ideally, if the solver allowed it, the range should be larger. However, giving larger ranges significantly increases the solving time.

	N1		N2		N3	
	param.	p-value	param.	p-value	param.	p-value
β_{COST}	-0.119	0.31	-0.185	0.34	-0.185	0.34
β_{TIME}	-0.301	0.57	-1.08	0.01	-1.08	0.01
$\bar{\mu}$	0.141	0.48	1.00	1.00	1.00	1.00
FLL	-165.5		-176.3		-176.3	

Table 4.4: Estimation results using the state-of-the-art continuous estimation of the different nesting structures when the ASCs are fixed to zero.

	Logit		Nested logit	
	lower bound	upper bound	lower bound	upper bound
ASC_{CAR}	-0.75	-0.25	-	-
ASC_{SM}	0.5	1	-	-
β_{COST}	-0.5	0	-0.5	0
β_{TIME}	-0.5	0	-0.5	0

Table 4.5: Upper and lower bounds of the parameters given to the MILP.

calculated as follows

$$e^R_{FLL} = 100 \left| \frac{FLL - \widehat{FLL}_R}{FLL} \right|. \tag{4.37}$$

	R	N1		N2		N3		Logit	
		FLL	e_{FLL} [%]	FLL	e_{FLL} [%]	FLL	e_{FLL} [%]	FLL	e_{FLL} [%]
MILP	5	-1648	958	-1560	866	-1558	860	-1344	729
	10	-358.1	130	-678.2	320	-369.8	128	-657.9	306
	20	-152.8	1.93	-180.9	12.1	-172.4	6.28	-160.1	1.29
	50	-153.7	1.32	-169.1	4.78	-171.2	5.54	-159.3	1.79
	100	-154.0	1.12	-168.6	4.46	-170.8	5.31	-161.0	0.757
Cont. estimation	-	-155.8	-	-161.4	-	-162.2	-	-162.2	-

Table 4.6: Final log-likelihood of the MILP by considering the optimal parameters from the continuous estimation.

As expected, the relative error decreases with the number of draws. For both the logit and for N1 the difference between the true FLL and the value obtained using the MILP is of less than 2% for 20 draws. For N3, the difference between the true FLL and the approximation using the MILP is a bit larger, but is also stable from 20 draws. For N2 the gap between the value obtained with the MILP and the value obtained with continuous estimation decreases from 12 % to 5 % when increasing the draws from 20 to 50.

We repeat the same as above for the values $\beta_{COST} = \beta_{TIME} = ASC_{SM} = ASC_{CAR} = 0$ and $\bar{\mu} = 0.5$. The results are summarized in Table 4.7. For N1, N3 and the logit, the relative error stabilizes, and is of less than 5 % from $R = 20$ draws. However, for nesting structure N2, the relative error of the FLL is of 60.4% for $R = 20$. This is due to the fact that under this configuration of parameters, there is one individual for which its chosen alternative is never the one with the highest value of the utility function in the simulated scenarios. Therefore, the contribution of this individual to the final log-likelihood is of $L_0 = -100$.

	R	N1		N2		N3		Logit	
		FLL	e_{FLL} [%]	FLL	e_{FLL} [%]	FLL	e_{FLL} [%]	FLL	e_{FLL} [%]
MILP	5	-1880	936	-3415	1520	-3316	1470	-1801	820
	10	-391.4	116	-639.0	204	-433.5	105	-417.1	113
	20	-185.1	2.00	-337.1	60.4	-226.7	7.19	-207.1	5.77
	50	-179.6	1.00	-224.1	6.61	-220.5	4.27	-199.5	1.89
	100	-179.5	1.07	-222.4	5.77	-216.6	2.44	-200.9	2.62
Cont. estimation	-	-181.5	-	-210.2	-	-211.4	-	-195.8	-

Table 4.7: Final log-likelihood of the MILP by considering the null model ($\beta = ASC = 0$; $\bar{\mu} = 0.5$).

These results justify using 20 draws for the MILP in the estimation process, based on a trade-off between the improvement in relative error and the computational time.

4.4.2 Estimating the logit model

To verify if the MILP framework can be used to correctly estimate the continuous parameters of a discrete choice model, we start by using it to estimate the parameters of the logit model corresponding to the specification of Table 4.2[15]. We do so using the MILP with 5, 10, 20 and 50 draws. We also introduce a stopping criteria, which is a time limit of 94h (338400s). Results are summarized in Table 4.8. We report the obtained parameters, together with the FLL, its relative error, the solution time, and the solution gap (in the cases where the time limit is reached). For 20 and 50 draws, the time limit is reached, so the problem is not solved to optimality. However, we see that the relative error of the final log-likelihood is already under 5%.

In order to evaluate the quality of the parameter estimation, we compute the relative error for each value of R and each parameter. Let β be the parameter obtained with the continuous estimation, and $\hat{\beta}_R$ be the result obtained with the MILP with R draws.

[15] All of the tested instances have been run in CPLEX Interactive Optimizer (CPLEX version 12.7.0.0) on a Unix server with 10 cores of 3.33 GHz and 62 GiB RAM.

	R	β_{TIME}	β_{COST}	ASC_{SM}	ASC_{CAR}	FLL	e_{FLL} [%]	time	gap [%]
MILP	5	-0.203	-0.0345	0.512	-0.255	-866.1	433	4 s	0
	10	-0.311	-0.480	0.500	-0.593	-361.8	123	4.6 min	0
	20	-0.414	-0.359	0.624	-0.749	-156.6	3.45	94 h	0.66
	50	-0.376	-0.478	0.832	-0.618	-157.5	2.89	94 h	7.6
Cont. estimation	-	-0.364	-0.376	0.778	-0.545	-162.2	-	-	-

Table 4.8: Results of the logit model

Then, we define the relative error as in Equation (4.37):

$$e_{\beta}^{R} = 100 \left| \frac{\beta - \hat{\beta}_R}{\beta} \right|. \tag{4.38}$$

Figure 4.3 shows the values of the relative errors for each of the parameters, as a function of the number of draws used. We can see that, in general, as the number of draws increases, the relative error decreases. This is not the case for 50 draws and β_{COST}, but it can be due to the fact that the problem is not solved to optimality. The relative errors of the parameters with 50 draws are of around 30% for β_{COST} and smaller for the rest. However, as discussed before, the relative error of the FLL is of less than 5 % already for 20 draws, meaning that this configurations of parameters also provide a good value of the final log-likelihood. This could be due to the fact that as shown in Table 4.3 the parameters β_{TIME} and β_{COST} have a p-value of 0.10 for the logit model.

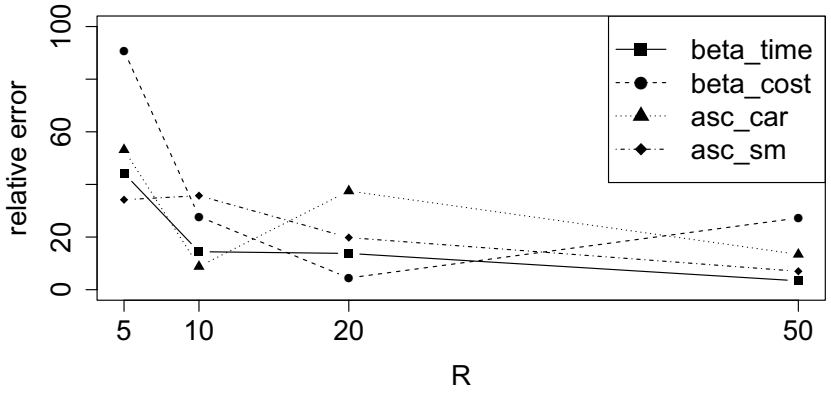


Figure 4.3: Relative errors as a function of the number of draws for each of the estimated parameters.

4.4.3 Estimating the nested logit model

We now estimate the continuous and the discrete parameters simultaneously. As for the logit model, we introduce a time limit of 48h (172800s) as a stopping criteria. The

obtained results are summarized in Table 4.9[16]. We report the values of the estimated continuous parameters, as well as the FLL, its relative error, the optimal nesting structure, the solution time and the solution gap (in the cases in which the time limit is reached). The optimal nesting structure is the correct one, independently of the number of draws used, which corresponds to N1 in Figure 4.2. Moreover, the relative error of the FLL is considerably low from $R = 10$. In fact, for $R = 10$ it is of less than 1%, which is probably due to a simulation artifact.

	R	β_{TIME}	β_{COST}	$\bar{\mu}$	FLL	e_{FLL} [%]	NS	time [h]	gap [%]
MILP	5	-0.315	-0.405	0.150	-1053	536	N1	1	0
	10	-0.375	-0.359	0.229	-165.3	0.165	N1	3.9	0
	20	-0.291	-0.265	0.135	-155.9	5.84	N1	48	17
Cont. estimation	-	-0.301	-0.119	0.141	-165.5	-	N1	-	-

Table 4.9: Results of the nested logit model

Next, we focus on the estimated parameters. Figure 4.4 shows the relative errors for each R and each parameter. The relative errors for β_{TIME} and $\bar{\mu}$ are low for $R = 5$. However, the $e^5_{\beta_{COST}}$ is of almost 250%. In this particular example, the parameter that is worse estimated is β_{COST}, but overall, the combined relative errors of the three parameters decrease as a function of the number of draws. As in the logit model case, the relative errors of the parameters are considerably large, but the value of the FLL is satisfying. This can be explained by looking back at Table 4.4. The p-values of β_{COST}, β_{TIME} and $\bar{\mu}$ are of 0.31, 0.57 and 0.48 respectively, meaning that changes in the values of the parameters do not necessarily imply big changes in the value of the final log-likelihood. In order to avoid this, and have more precise values of the parameter estimates, we should consider a larger number of respondents.

In this case study we have shown that the simulation error of final log-likelihood with the proposed framework is relatively small (of around 5%) from R=20 draws. We have shown that the discrete parameters are estimated correctly (i.e: the optimal nesting structure is identified by using the MILP). However, there are several limitations due to the complexity of the problem. First, the maximum number of respondents that we can consider in order to solve the problem is $N = 200$. This is not enough to have significant parameters at the 10% level with the continuous estimation software, and therefore the parameter estimates we obtain with the MILP are not accurate, even if their relative error decreases as a function of the number of draws R. Second, the maximum number of draws that we can consider is of $N = 50$ for the estimation of the parameters of a logit model, and of $N = 20$ for the estimation of the discrete and continuous parameters of the nested logit model. If we consider a higher number of draws, the server that

[16] All of the tested instances have been run in CPLEX Interactive Optimizer (CPLEX version 12.7.0.0) on a Unix server with 10 cores of 3.33 GHz and 62 GiB RAM.

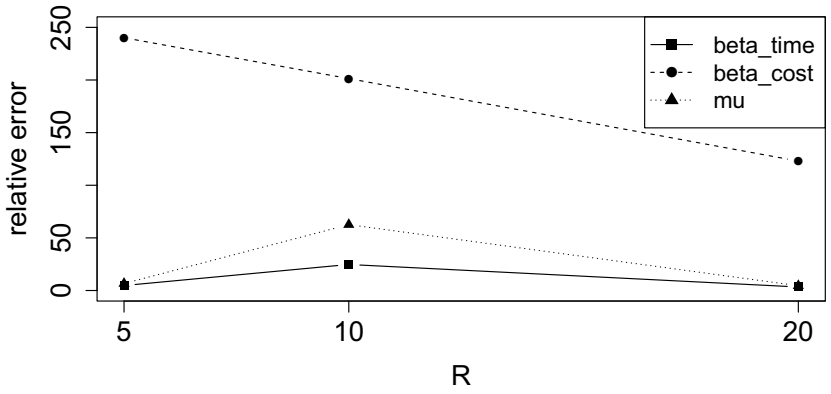


Figure 4.4: Relative errors as a function of the number of draws for each of the estimated parameters.

we are using to solve the problem runs out of memory, due to the complexity of the problem. Moreover, both for the logit and for the nested logit, optimality is not proven by the standard solver that we use. Third, the number of parameters that we are able to estimate is limited, due to Constraint (A.5). For the logit model, we are able to estimate β_{COST}, β_{TIME}, ASC_{SM} and ASC_{CAR}, but for the estimation of the nested logit model, we must normalize both ASCs to zero.

All of the limitations discussed above are due to the computational complexity of the proposed framework, and not to the framework itself. There are three directions of research to investigate: (i) to tighten the formulation, (ii) to use dedicated algorithms that exploit the structure of the problem, and (iii) to use heuristics instead of exact methods.

4.5 Conclusions and future work

We have introduced the concept of *discrete-continuous maximum likelihood* and shown that it can be modeled as a mixed integer linear program. This framework allows to simultaneously estimate the (continuous) parameters of the utility function as well as discrete parameters. In this chapter, the discrete parameters considered are the allocation of alternatives to nests. Classically, the values of the latter are given as an input, and in this chapter we estimate them from data.

Moreover, the proposed formulation is general in the sense that, by changing the assumption on the distribution of the error terms, any type of discrete choice model can be estimated (conditional to the fact that the draws can be generated). For instance, the framework that has been presented is straightforward to adapt to error component

models. Since we have found a linear approximation of the log-likelihood function, global optimality can be insured. This might be useful in discrete choice models when the exact expression of the log-likelihood is non-concave, such as latent class models.

We have considered as an illustrative example a case study with three alternatives so that a full enumeration of the nesting structures is possible, in order to have a benchmark for the results that we obtain.

Results with the logit model show that the MILP provides relatively good results for the value of the final log-likelihood, and that a minimum of 20 draws is needed to obtain reliable results in this case study. By applying the MILP framework to the problem of discrete-continuous maximum likelihood, we find that the identified nesting structure is the correct one, and that the relative error of the final log-likelihood is of around 5% from 20 draws. However, the obtained values with the MILP of the continuous parameters (both for the logit and the nested logit) are not precise, as their relative errors are considerably large. This is probably related to the relatively small number of individuals that we are using for the estimation.

As the framework has been presented, standard errors of the parameters are not reported. To address this, bootstrapping could be used. By doing so, we would obtain the empirical distribution of the estimated parameters, and we would be able to calculate its standard errors. Bootstrapping would also reveal if the estimated discrete parameters (the optimal nesting structure in our case) is robust or not.

The goal of this case study is to show that the proposed MILP is able to solve the discrete-continuous maximum likelihood problem. However, to make it operational in practice, the number of individuals, draws and alternatives must be larger than what has been shown in this chapter. To do so, the solving time must be decreased. The discrete-choice literature must borrow from combinatorial optimization to do so. There are three directions to investigate. First, the MILP specification can be improved by using valid inequalities, or other modeling techniques to tighten the formulation. Second, the solution algorithm can also be improved. In this chapter, we have used a standard solver. Dedicated algorithms, exploiting the structure of the problem by proper decomposition techniques, should be used instead. Techniques such as column generation or Lagrangian relaxation, should be implemented, since the problem is separable for each of the possible nesting structures. Finally, if after tightening the problem and using dedicated algorithms the problem cannot be solved exactly for realistic sizes, heuristics could also be investigated.

Once it is possible to solve realistic sizes of the problem, this framework opens the door to new types of research. For instance, the membership of an alternative to the choice set of an individual could be added to the estimation process by means of discrete parameters. The framework could also be used to determine the model specification

(i.e.: the expression of the utilities). In other words, the functional form in which variables enter the utilities could also be part of the estimation process, by adding several functional forms that multiply mutually exclusive discrete parameters. In summary, the tools that have been introduced in this chapter are thought to be a first step towards more automatic and data-driven discrete decisions, that are currently taken on a trial-and-error basis by the modeler.

5

Conclusion

In this chapter, we review the main findings of this thesis, as well as its theoretical and practical implications (Section 5.1), and finalize by outlining future research directions in Section 5.2.

5.1 Main findings and implications

This thesis proposes methods to address correlations in discrete choice models. We have addressed two types of correlations: correlations within alternatives (endogeneity) and correlations between alternatives. Addressing these aspects is crucial to obtain correct demand indicators and has motivated the development of the methods presented. Moreover, we have applied these methods to real case studies to show their applicability, and to gain insights in mode choice and purchase of private motorized modes.

Chapter 2 addresses the correlation between observed and unobserved attributes of an alternative. We have focused on the novel method introduced by Guevara and Polanco (2016), the multiple indicator solution. We have shown, from a theoretical point of view, that it can also be used when there are interactions between the observed and the unobserved attributes. We have shown that it is operational, by applying it to a mode choice case study of revealed preference data in Switzerland. To show that the methodology is useful, we have compared the results obtained with a state-of-the-art solution: the integrated choice and latent variable model.

In Chapter 3 we have addressed two main challenges. First, we have dealt with revealed preference data when information is only available for the chosen alternative. We have had to (i) define the choice set for each respondent, and (ii) impute the attributes of the non-chosen alternatives. To define the choice set, we have defined an alternative as a combination of a market segment and a fuel type, resulting in 15 different alternatives. Then, to impute the attributes of the unchosen alternatives, we have used multiple imputations from the empirical distributions of these attributes. Second, we have used a cross nested logit to take into account that similar alternatives share unobserved attributes. The nesting structure is characterized by market segment and by fuel type. In other words, we have assumed that alternatives sharing either market segment or fuel type share unobserved characteristics, and therefore belong the the same nest.

Motivated by the need to decide the best nesting structure in Chapter 3, Chapter 4 proposes a way to simultaneously estimate discrete and continuous parameters using maximum likelihood. We have used it to determine the continuous parameters of the utilities together with the discrete parameters that characterize the best nesting structure. To do so, we have introduced the concept of *discrete-continuous maximum likelihood*. We have linearized the log-likelihood estimator and formulated the problem as a mixed integer linear problem. The framework is easy to generalize to any model where the error terms are easy to simulate. This can be used to estimate the parameters of models for which the log-likelihood is non-concave, since for an MILP, algorithms that can find the global optimum exist. In this chapter, we have introduced the DCML framework, and shown under which circumstances the problem is computationally feasible.

The benefits of the methods described above have been shown in real case studies related to transportation. We have gained insights in private motorized modes, both in terms of modal split and the car market itself, as described below. It is important to note, however, that they can be easily extended to any other domain where discrete choice models are applied (social welfare, tax distribution, health economics, evacuation decisions, environmental economics, and marketing among others).

To gain insights on modal split, we have used the PostBus case study (Chapter 2). We have shown that endogeneity was present in the model, and corrected for it. We have derived the value of time of individuals as a function of their *car loving* attitude and their income, and found values that are larger than what is reported in the literature. However, the values found in the literature depend only on the trip purpose, but not on the preference towards car, nor on the income level of the respondents. We have also calculated the travel time elasticities and have found that the logit model underestimates it. In conclusion, it is necessary to correct for endogeneity in order to obtain accurate demand indicators.

We have also studied the new car market (Chapter 3), in particular for hybrid and electric vehicles. We have shown that a cross nested logit model is more adequate to model the

car-type choice (due to the nature of the definition of the alternatives) and derived elasticities as well as willingness to pay indicators, and market shares in different future policy scenarios. The values of the elasticities are in line with those from the literature, and the values of the willingness to pay are in line with the real market of new cars. We have also found that the most effective scenario to increase the sales of electric vehicles corresponds to the *technological innovation* scenario (among the scenarios that we have defined) and that the new sales of new cars under this scenario would be of around 1%.

In summary, from a theoretical point of view, we have added to the field of discrete choice by (i) extending existing methods, namely, the MIS method, and (ii) proposing new methods, namely, the imputation of the attributes of the non chosen alternatives, and the discrete-continuous maximum likelihood framework. These new tools are interesting for both researchers, since they open the door to unexplored research directions, and for practitioners. Practitioners have now new methods to use revealed preference data and to obtain accurate estimates and demand indicators. This is particularly interesting nowadays, since data availability is increasing, but it is more and more often not collected for the purpose of applying discrete choice models.

5.2 Future research directions

Correlations between and within alternatives have been treated separately in this thesis. However, they often happen simultaneously, and this should be taken into account. For example, in Chapter 3, addressing price endogeneity might have an impact in the demand indicators produced by the model. This should be investigated further, and could be done using the methodology introduced in Chapter 2. Another possibility, is to consider attitudes and perceptions by means of the ICLV. Preliminary results (not included in the thesis) show that the *car loving* attitude plays a role in the purchases of new cars. It would be interesting to study this further. In conclusion, it is important that both types of correlations are dealt with more often in the literature, or at least considered.

Related to *discrete-continuous maximum likelihood*, and to the MILP formulation of the maximum likelihood problem, there is still a lot of work to be done. The model needs to be tested in other case studies where the number of alternatives and nesting structures is larger. However, before doing so, the model must be faster to solve. Future research directions include exact methods (lagrangian relaxation, column generation) and heuristics. Also, we claim that the model is applicable to other choice models, such as latent class and error component, but this still needs to be tested. It would be interesting to show all the models that can be estimated using this framework. Last, as the framework is now, standard errors are not reported. An option, is to use bootstrapping to calculate the standard errors. The MILP can be run several times with different draws, and the

empirical distribution of the estimated parameters will be obtained, from which the standard errors can be computed. It would also be interesting to identify discrete variables (other than nest allocation ones) in which our framework could be used. Some examples might include the membership to the choice set, the specification of the utilities, or if endogeneity is present or not.

In this thesis, we have focused on modeling better the error terms, either by modeling explicitly unobserved characteristics, or by working on the unobserved correlation between alternatives. The discussions on the model specification have been limited. There is always the implicit assumption that a model is correctly specified, which is not necessarily the case. It would therefore be desirable to study further the model specification of the different models that have been introduced.

All the case studies that have been addressed in this thesis correspond to static data. They represent the state of the market at one moment in time. Both for mode choice and for the new car market, it would be very interesting to study the dynamics over time of people's decisions. Some work in this direction has been done by Glerum (2014). It would be interesting to include the dynamics in our case studies, but the data needed to do so is not yet available.

Data availability is increasing exponentially nowadays, thanks to new data sources, such as mobile applications. It is important that the field of discrete choice models learns how to take advantage of them. This type of data is not straight forward to use, for several reasons. First, its quality is often not good, in the sense that there are missing values, and we only have information on the chosen alternatives. A first step in this direction is to define the consideration choice set as well as the unchosen alternatives, as we have done in Chapter 3. It would be interesting to apply the multiple imputations solution that we propose in a dataset that has not been collected for the purpose of DCM. How to deal with missing values also needs to be investigated, but multiple imputations could also be an option. Second, there are privacy and legal issues, linked with the ethical use of the data. Is it moral to use the data from an application to maximize the revenue of a company, even when people are not aware that this data is being collected? There needs to be a social debate before this new type of data can become widely used. It is however undeniable that the data revolution is taking place, and that companies are becoming more interested in data-driven approaches in general, and in discrete choice modeling in particular, for pricing, maximizing revenue, deriving elasticities and performing market segmentation.

A

Linearization of the expression of the probabilities

This appendix is a summary of the part of Pacheco et al. (2017) that we borrow to develop the methodology presented in Chapter 4. In particular, we use the idea of using simulation to dispose of the nonlinearities caused by the expression of the probabilities.

From the definition of w_{inr}

$$w_{inr} = \begin{cases} 1 & \text{if} \quad U_{inr} > U_{jnr} \quad \forall j \neq i, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall i \in \mathcal{C}_n, n, r, \tag{A.1}$$

the chosen alternative of an individual n in a scenario r corresponds to its associated highest utility (U_{nr}). We introduce U_{nr} that is defined as

$$U_{nr} = \max_{i \in \mathcal{C}_n} U_{inr}, \qquad \forall n, r. \tag{A.2}$$

In order to linearize this expression, we must define lower and upper bounds of U_{inr}. We note them as ℓ_{inr} and m_{inr}, respectively. Then,

$$\ell_{inr} \leq U_{inr} \leq m_{inr}, \qquad \forall i, n, r. \tag{A.3}$$

We remind the reader that in our framework, $U_{in} = V_{in} + \varepsilon_{in}$, with $V_{in} = \sum_k \beta_k x_{in}^k$. Since the values of x_{in}^k are given, we must impose upper and lower bounds of β_k to insure the existence of ℓ_{inr} and m_{inr}[17]

[17] Note that in Pacheco et al. (2017) it is the opposite: the values of the attributes are decision variables

The following expression is used to linearize Equations (A.1) and (A.2) (see Pacheco et al. (2017) for the proof)

$$U_{inr} \leq U_{nr}, \qquad \forall i \in \mathcal{C}_n, n, r, \tag{A.4}$$

$$U_{nr} \leq U_{inr} + M_{nr}(1 - w_{inr}), \qquad \forall i \in \mathcal{C}_n, n, r, \tag{A.5}$$

where

- $m_{nr} = \max_{j \in \mathcal{C}_n} m_{jnr}$ is the largest upper bound across all alternatives,
- $\ell_{nr} = \min_{j \in \mathcal{C}_n} \ell_{jnr}$ is the smallest lower bound across all alternatives, and
- $M_{nr} = m_{nr} - \ell_{nr}$ is the difference between the largest upper bound and the smallest lower bound.

Note that the value of M_{nr} depends of the upper and lower bounds ℓ_{inr} and m_{inr}, that depend on the bounds of β_k. Therefore, the tighter the bounds on β_k, the tighter the MILP formulation.

Other constraints so that w_{inr} are well defined defined. Since only available alternatives can be chosen by an individual, we add the following constraint

$$w_{inr} \leq y_{in}, \qquad \forall i, n, r. \tag{A.6}$$

Finally, as each individual chooses exactly one alternative in each scenario, we impose

$$\sum_{i=1}^{J} w_{inr} = 1, \qquad \forall n, r. \tag{A.7}$$

in their model, and the β_ks are considered to be known.

B

An MILP formulation for the error component model

The framework that has been introduced in Section 4.3.2 to model nested logit models is straightforward to adapt to error component models. This appendix explains how.

Equation (4.14) is changed by:

$$\varepsilon_{in} = \omega_{im} + \nu_{in} \tag{B.1}$$

where:

- $\omega_{im} \overset{iid}{\sim} N(0, \sigma_m^2)$
- $\nu_{in} \overset{iid}{\sim} \text{EV}(0, 1)$

From the properties of the normal distribution, we know that if $\omega_{im} \overset{iid}{\sim} N(0, \sigma_m^2)$, then

$$\omega_{im} = \sigma_m \omega'_{im}, \tag{B.2}$$

verifies that $\omega'_{im} \overset{iid}{\sim} N(0, 1)$. Note that the lower bound of σ_m can be defined as 0, since the variance of the normal random variable is its square. Then, the equivalent to

Equation (4.18) is now:

$$U_{in} = V_{in} + \sigma_m \omega'_{im} + \nu_{in}. \tag{B.3}$$

By considering the definition of b_{im} as in Equation (4.19), we can express the utility function as follows

$$U_{in} = V_{in} + \nu_{in} + \sum_{m=1}^{M} b_{im}\sigma_m\omega_{im}. \tag{B.4}$$

By defining $\kappa_{im} = b_{im}\sigma_m$, the nonlinearity produced by $b_{im}\sigma_m$ can be linearized by the following constraints

$$\kappa_{im} \leq u_m b_{im} \qquad \forall i, m \tag{B.5}$$
$$\kappa_{im} \leq \sigma_m \qquad \forall i, m \tag{B.6}$$
$$\kappa_{im} \geq \sigma_m - u_m(1 - b_{im}) \qquad \forall i, m \tag{B.7}$$
$$\kappa_{im} \geq 0 \qquad \forall i, m \tag{B.8}$$

where u_m is an upper bound of the value of σ_m. We have that $\sigma_m \in [0, u_m]$.

The constraints introduced in Section 4.3.2 so that each alternative belongs to exactly one nest (Constraint (4.20)), the symmetry breaking constraints (Constraints (4.36)), and the constraints related to the maximum number of alternatives per nest (Constraint (4.35)) remain unchanged, since the interpretation of b_{im} is analogous.

Finally, for identification purposes we also need a constraint equivalent to (4.30). In the case of the error component model, this constraint is as follows

$$\text{if} \quad \sum_{i=1}^{J} b_{im} \leq 1 \quad \text{then} \quad \sigma_m = 0, \qquad \forall m. \tag{B.9}$$

To linearize this implication, we introduce binary variables t_m that take value 0 if $\sum_{i=1}^{J} b_{im} \leq 1$. The linearization is then as follows

$$\sum_{i=1}^{J} b_{im} \geq 2t_m, \qquad \forall m, \tag{B.10}$$
$$\sigma_m \leq t_m u_m, \qquad \forall m. \tag{B.11}$$

To prove the equivalence we consider the following:

- If $\sum_{i=1}^{J} b_{im} = 0$, constraints (B.10) become

$$0 \geq 2t_m, \qquad \forall m, \tag{B.12}$$

  which is equivalent to $t_m \leq 0$, and since it is a binary variable, we obtain that $t_m = 0$. Then, from Equation (B.11) we have that $\sigma_m \leq 0$. Since by definition, $\sigma_m \in [0, u_b]$, we obtain that $\sigma_m = 0$.

- If $\sum_{i=1}^{J} b_{im} = 1$, constraints (B.10) become

$$1 \geq 2t_m, \qquad \forall m, \tag{B.13}$$

  which is equivalent to $t_m \leq 0.5$, and since it is a binary variable, we obtain that $t_m = 0$. Analogously, we obtain that $\sigma_m = 0$.

- If $\sum_{i=1}^{J} b_{im} \geq 2$, Equation (B.10) is always true (both with $t_m = 0$ and with $t_m = 1$), therefore t_m is free, and Equation (B.11) is always verified.

C

Previous formulations of the MILP

The MILP model presented in Chapter 4 is the final result of a long modeling exercise. In this appendix we show the preliminary versions of the linearization of the log-likelihood function. We show it for the logit model, since it is simpler, but it is straight forward to adapt it to the nested logit case. What differs from one model to the next is the way we linearized the log function from Equation (4.7). For convenience, we repeat the equation here:

$$\sum_{n=1}^{N} \sum_{i \in \mathcal{C}_n} d_{in} \left(\log \left(\sum_{r=1}^{R} w_{inr} \right) - \log(R) \right). \tag{C.1}$$

C.1 Logit model 1

In our attempt to linearize the Expression (C.1), our first idea came from the realization that the argument of the logarithm can only take integer values from 0 to R and that it is possible to precompute the logarithm for each of these values. Then, we can linearize it by introducing binary variables denoted γ_{inp} defined as follows

$$\gamma_{inp} = \begin{cases} 1 & \text{if } \sum_{r=1}^{R} w_{inr} = p, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall i, n, p. \tag{C.2}$$

Then, Equation (C.1) is equivalent to

$$\sum_{n=1}^{N}\sum_{i=1}^{I} d_{in}\left(\sum_{p=0}^{R}\gamma_{inp}L_p - \log(R)\right), \tag{C.3}$$

where $L_p = \log(p)$, $p = 1, ..., R$ and $L_0 = -100$ is a pre-processed vector of $R+1$ components.

Linearization of γ_{inp} Equation (C.2) can be linearized as follows

$$(R-p+1)\delta^1_{inp} - 1 \geq \sum_{r=1}^{R} w_{inr} - p, \qquad \forall i, n, p, \tag{C.4}$$

$$(p+1)\delta^2_{inp} - 1 \geq p - \sum_{r=1}^{R} w_{inr}, \qquad \forall i, n, p, \tag{C.5}$$

$$\delta^1_{inp} + \delta^2_{inp} - 2\gamma_{inp} \leq 1, \qquad \forall i, n, p, \tag{C.6}$$

$$\sum_{p=1}^{R}\gamma_{inp} = 1, \qquad \forall i, n, \tag{C.7}$$

where $\delta^1_{inp}, \delta^2_{inp}$ are binary variables. To prove the equivalence between Equation (C.2) and Equations (C.4)-(C.7) we consider three cases:

- If $\sum_{r=1}^{R} w_{inr} = p$, constraints (C.4)-(C.5) become

$$(R-p+1)\delta^1_{inp} - 1 \geq 0 \qquad \forall i, n, p, \tag{C.8}$$

$$(p+1)\delta^2_{inp} - 1 \geq 0 \qquad \forall i, n, p. \tag{C.9}$$

  Constraints (C.8) and (C.9) impose that $\delta^1_{inp} = \delta^2_{inp} = 1$. Using this, constraint (C.6) is written

$$2 - 2\gamma_{inp} \leq 1 \iff 1 \leq 2\gamma_{inp} \iff \gamma_{inp} = 1. \tag{C.10}$$

  From constraint (C.7), $\gamma_{inr} = 0$ if $r \neq p$.

- If $\sum_{r=1}^{R} w_{inr} > p$, constraint (C.4) becomes

$$(R-p+1)\delta^1_{inp} - 1 \geq \sum_{r=1}^{R} w_{inr} - p > 0 \iff (R-p+1)\delta^1_{inp} > 1 \iff \delta^1_{inp} = 1 \tag{C.11}$$

  From constraint (C.7) we obtain that $\gamma_{inp} = 0$, so from constraint (C.6) $\delta^2_{inp} = 0$ and constraint (C.6) is trivial.

- If $\sum_{r=1}^{R} w_{inr} < p$, the derivation is analogous to the previous case.

Therefore the MILP maximum log-likelihood problem can be formalized as follows:

$$\max \quad \sum_{n=1}^{N}\sum_{i=1}^{I} d_{in}\left(\sum_{p=0}^{R}\gamma_{inp}L_p - \log R\right)$$

$$\begin{aligned}
\text{subject to} \quad & U_{inr} = V_{in} + \tilde{\varepsilon_{inr}} && \forall i \in \mathcal{C}_n, n, r && \text{(C.12)}\\
& w_{inr} \leq y_{in} && \forall i, n, r && \text{(C.13)}\\
& U_{inr} \leq U_{nr} && \forall i \in \mathcal{C}_n, n, r && \text{(C.14)}\\
& U_{nr} \leq U_{inr} + M_{nr}(1 - w_{inr}) && \forall i \in \mathcal{C}_n, n, r && \text{(C.15)}\\
& \sum_{i=1}^{I} w_{inr} = 1 && \forall n, r && \text{(C.16)}\\
& (R - p + 1)\delta^1_{inp} - 1 \geq \sum_{r=1}^{R} w_{inr} - p && \forall i, n, p && \text{(C.17)}\\
& (p+1)\delta^2_{inp} - 1 \geq p - \sum_{r=1}^{R-1} w_{inr} && \forall i, n, p && \text{(C.18)}\\
& \delta^1_{inp} + \delta^2_{inp} - 2\gamma_{inp} \leq 1 && \forall i, n, p && \text{(C.19)}\\
& \sum_{p=0}^{R} \gamma_{inp} = 1 && \forall i, n && \text{(C.20)}\\
& w_{inr}, \delta^1_{inr}, \delta^2 inr \in \{0, 1\} && \forall i, n, r && \text{(C.21)}\\
& \gamma_{inp} \in \{0, 1\} && \forall i, n, p && \text{(C.22)}\\
& \beta \in R^s && && \text{(C.23)}\\
& U_{inr} \in R && \forall i, n, r && \text{(C.24)}\\
& U_{nr} \in R && \forall n, r && \text{(C.25)}
\end{aligned}$$

where

- s is the number of estimated parameters,
- $M_{nr} = u_{nr} - \ell_{nr}$,
- $u_{nr} = \max_{i \in \mathcal{C}_n} U_{inr}$,
- $\ell_{nr} = \min_{i \in \mathcal{C}_n} U_{inr}$,
- y_{in} are the observed availabilities,
- d_{in} are the observed choices,

This formulation contains three big M constraints (constraints (C.15), (C.17) and (C.18)), as well as $3\text{x}I\text{x}N\text{x}P$ extra binary variables (γ,δ^1 and δ^2), which is costly in terms of computation time.

C.2 Logit model 2: ordering gamma

An improvement to the previous model comes from the idea of ordering the γ variables of the previous section. We define Ω_{inr} as

$$\text{if} \quad \sum_{r=1}^{R} w_{inr} = p \implies \begin{cases} \Omega_{ink} = 1 & \forall k \leq p \\ \Omega_{ink} = 0 & \forall k > p \end{cases} \qquad \forall i, n, p. \tag{C.26}$$

This can be linearized as follows

$$(R - p + 1)\Omega_{inp} - 1 \geq \sum_{r=1}^{R} w_{inr} - p \qquad \forall i, n, p \tag{C.27}$$

$$\sum_{r=1}^{R} w_{inr} = \sum_{r=1}^{R} \Omega_{inr} \qquad \forall i, n \tag{C.28}$$

The proof of this equivalence is straight forward from the previous section. The discrete-continuous maximum likelihood problem can then be formalized as

$$\max \quad \sum_{n=1}^{N}\sum_{i=1}^{I} d_{in}\left(\sum_{r=1}^{R}\Omega_{inr}\Delta L_p - \log R\right)$$

$$\text{subject to} \quad U_{inr} = V_{in} + \tilde{\varepsilon_{inr}} \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.29)}$$

$$w_{inr} \leq y_{in} \qquad \forall i, n, r \qquad \text{(C.30)}$$

$$U_{inr} \leq U_{nr} \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.31)}$$

$$U_{nr} \leq U_{inr} + M_{nr}(1 - w_{inr}) \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.32)}$$

$$\sum_{i=1}^{I} w_{inr} = 1 \qquad \forall n, r \qquad \text{(C.33)}$$

$$(R - p + 1)\Omega_{inp} - 1 \geq \sum_{r=0}^{R-1} w_{inr} - p \qquad \forall i, n, p \qquad \text{(C.34)}$$

$$\sum_{r=1}^{R}\Omega_{inr} = \sum_{r=1}^{R} w_{inr} \qquad \forall i, n \qquad \text{(C.35)}$$

$$w_{inr}, \Omega_{inr} \in \{0, 1\} \qquad \forall i, n, r \qquad \text{(C.36)}$$

$$\beta \in R^s \qquad \text{(C.37)}$$

$$U_{inr} \in R \qquad \forall i, n, r \qquad \text{(C.38)}$$

$$U_{nr} \in R \qquad \forall n, r \qquad \text{(C.39)}$$

Where

- $\Delta L_p = \log(p + 1) - \log(p), \forall p \geq 1,$
- $\Delta L_0 = -100$

and the rest of the notations are the same as in Section C.1.

We can see that this formulation has one less big M constraint compared to the previous formulation, and 2xIxNxR less binary variables. The main disadvantage of this formulation is that it has many symmetries, given that there are many ways to order Ω_{inr} since it is a binary variable.

C.3 Logit model 3: assignment problem for the ordering of gamma

Inspired from the previous formulation, the ordering problem can also be seen as an assignment problem. We can introduce δ_{inlk} binary variables such that δ_{inlk} takes value

1 if the value of w_{inl} is assigned to Ω_{inl}, and 0 otherwise, as follows

$$\delta_{inlk} = \begin{cases} 1 & \text{if} \quad \Omega_{ink} \leftarrow w_{inl} \\ 0 & \text{otherwise} \end{cases} \qquad \forall i, n, l, k, \tag{C.40}$$

Then, as in the classical assignment problem, Ω_{inl} can be expressed as

$$\Omega_{inl} = \sum_{k=1}^{R} \delta_{inlk} w_{ink}, \qquad \forall i, n, l. \tag{C.41}$$

Since the product of δ_{inlk} and w_{ink} has to be linearized, we introduce variables $t_{inkl} = \delta_{inlk} w_{ink}$. Then, by adding

$$\Omega_{ink} \geq \Omega_{in(k+1)}, \forall i, n, k, \tag{C.42}$$

the assignment problem orders the Ω variables in decreasing order.

This can be linearized as follows:

$$\sum_{l=1}^{R} \delta_{inlk} = 1 \qquad \forall i, n, k \tag{C.43}$$

$$\sum_{k=1}^{R} \delta_{inkl} = 1 \qquad \forall i, n, l \tag{C.44}$$

$$t_{inlk} \leq \delta_{inlk} \qquad \forall i, n, l, k \tag{C.45}$$

$$t_{inkl} \leq w_{ink} \qquad \forall i, n, l, k \tag{C.46}$$

$$t_{inlk} \geq \delta_{inlk} + w_{ink} - 1 \qquad \forall i, n, l, k \tag{C.47}$$

Where Constraints (C.43) and (C.44) are the classical constraints of the assignment problem, and constraints (C.45)-(C.47) are for the linearization of $\delta_{inlk} w_{ink}$.

The complete MILP is then

$$\max \quad \sum_{n=1}^{N}\sum_{i=1}^{I} d_{in}\left(\sum_{r=1}^{R}\Omega_{inr}\Delta L_p - \log R\right)$$

subject to

$$U_{inr} = V_{in} + \tilde{\varepsilon_{inr}} \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.48)}$$

$$w_{inr} \leq y_{in} \qquad \forall i, n, r \qquad \text{(C.49)}$$

$$U_{inr} \leq U_{nr} \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.50)}$$

$$U_{nr} \leq U_{inr} + M_{nr}(1 - w_{inr}) \qquad \forall i \in \mathcal{C}_n, n, r \qquad \text{(C.51)}$$

$$\sum_{i=1}^{I} w_{inr} = 1 \qquad \forall n, r \qquad \text{(C.52)}$$

$$\sum_{k=1}^{R} \delta_{inkl} = 1 \qquad \forall i, n, l \qquad \text{(C.53)}$$

$$\sum_{l=1}^{R} \delta_{inlk} = 1 \qquad \forall i, n, k \qquad \text{(C.54)}$$

$$\sum_{k=0}^{R-1} t_{inlk} = \Omega_{inl} \qquad \forall i, n, l \qquad \text{(C.55)}$$

$$t_{inlk} \leq \delta_{inkl} \qquad \forall i, n, l, k \qquad \text{(C.56)}$$

$$t_{inkl} \leq w_{ink} \qquad \forall i, n, l, k \qquad \text{(C.57)}$$

$$\delta_{inlk} + w_{ink} - 1 \leq t_{inkl} \qquad \forall i, n, l, k \qquad \text{(C.58)}$$

$$\Omega_{ink} \leq \Omega_{in(k-1)} \qquad \forall i, n, l, k \qquad \text{(C.59)}$$

$$w_{inr}, \Omega_{inr} \in \{0, 1\} \qquad \forall i, n, r \qquad \text{(C.60)}$$

$$\delta_{inkl}, t_{inlk} \in \{0, 1\} \qquad \forall i, n, l, k \qquad \text{(C.61)}$$

$$\beta \in R^s \qquad \text{(C.62)}$$

$$U_{inr} \in R \qquad \forall i, n, r \qquad \text{(C.63)}$$

$$U_{nr} \in R \qquad \forall n, r \qquad \text{(C.64)}$$

The advantage of this model is that there is only one big M constraint. However, this comes at the price of many binary variables. The motivation of trying this formulation is that in the assignment problem, the solution of the relaxation problem is directly the solution of the MILP (Bierlaire, 2015). However, this model was proven to be slower to solve than the one presented in Section C.2. We think this is due to the fact that in this case, the assignment problem assigns decision variables to other decision variables, making it slower than the classical assignment problem.

Bibliography

Abbe, E., Bierlaire, M., and Toledo, T. (2007). Normalization and correlation of cross-nested logit models, *Transportation Research Part B: Methodological* **41**(7): 795–808.

Abou-Zeid, M., Ben-Akiva, M., Bierlaire, M., Choudhury, C. and Hess, S. (2010). Attitudes and value of time heterogeneity, *in* E. V. de Voorde and T. Vanelslander (eds), *Applied Transport Economics A Management and Policy Perspective*, de boeck, pp. 523–545.

Adda, J. and Cooper, R. (2000). Balladurette and Juppette: A Discrete Analysis of Scrapping Subsidies, *Journal of Political Economy* **108**(4): 778–806.

Adepetu, A. and Keshav, S. (2017). The relative importance of price and driving range on electric vehicle adoption: Los Angeles case study, *Transportation* **44**(2): 353–373.

Ajzen, I. (2001). Nature and Operation of Attitudes, *Annual Review of Psychology* **52**(1): 27–58.

Amemiya, T. (1978). The estimation of a simultaneous equation generalized probit model, *Econometrica* **46**(5): 1193–1205.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association* **91**(434): 444–455.

Anowar, S., Eluru, N. and Miranda-Moreno, L. F. (2014). Alternative Modeling Approaches Used for Examining Automobile Ownership: A Comprehensive Review, *Transport Reviews* **34**(4): 441–473.

Atasoy, B., Glerum, A. and Bierlaire, M. (2013). Attitudes towards mode choice in Switzerland, *disP - The Planning Review* **49**(2): 101–117.

Autobild (2017). Audi rs3 sportback, `http://http://www.autobild.es/coches/audi/a3/rs3-sportback-5-2015`. Accessed on 12.06.2017.

Axhausen, K. W., Hess, S., König, A., Abay, G., Bates, J. J. and Bierlaire, M. (2008). Income and distance elasticities of values of travel time savings: New Swiss results, *Transport Policy* **15**(3): 173–185.

Bart, Y., Stephen, A. T. and Sarvary, M. (2014). Which Products Are Best Suited to Mobile Advertising? A Field Study of Mobile Display Advertising Effects on Consumer Attitudes and Intentions, *Journal of Marketing Research* **51**(3): 270–285.

Baum, C. F. (2006). *An Introduction to Modern Econometrics Using Stata*, Stata Press.

Beck, M. J., Rose, J. M. and Greaves, S. P. (2017). I can't believe your attitude: a joint estimation of best worst attitudes and electric vehicle choice, *Transportation* **44**(4): 753–772.

Beck, M. J., Rose, J. M. and Hensher, D. A. (2013). Environmental attitudes and emissions charging: An example of policy implications for vehicle choice, *Transportation Research Part A: Policy and Practice* **50**: 171–182.

Ben-Akiva, M. and Boccara, B. (1987). Integrated framework for travel behavior analysis, *International Association of Travel Behavior Research (IATBR) Conference, Aix-en-Provence, France*.

Ben-Akiva, M., Bolduc, D. and Bradley, M. (1993). Estimation of travel choice models with randomly distributed values of time, *Transportation Research Record* (1413).

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press series in transportation studies, MIT Press.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., Palma, A. D., Gopinath, D., Karlstrom, A. and Munizaga, M. A. (2002). Hybrid Choice Models: Progress and Challenges, *Marketing Letters* **13**(3): 163–175.

Berkovec, J. (1985). Forecasting automobile demand using disaggregate choice models, *Transportation Research Part B: Methodological* **19**(4): 315–329.

Berkovec, J. and Rust, J. (1985). A nested logit model of automobile holdings for one vehicle households, *Transportation Research Part B: Methodological* **19**(4): 275–285.

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile Prices in Market Equilibrium, *Econometrica* **63**(4): 841–890.

Berry, S., Levinsohn, J. and Pakes, A. (1998). Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market, *Technical Report w6481*, National Bureau of Economic Research, Cambridge, MA.

Berry, S., Linton, O. B. and Pakes, A. (2004). Limit theorems for estimating the parameters of differentiated product demand systems, *The Review of Economic Studies* **71**(3): 613–654.

Bierlaire, M. (2006). A theoretical analysis of the cross-nested logit model, *Annals of Operations Research* **144**(1): 287–300.

Bierlaire, M. (2015). *Optimization: Principles and Algorithms*, EPFL Press, Lausanne.

Bierlaire, M. (2016). Pythonbiogeme: a short introduction, *Technical Report TRANSP-OR 160706, Series on Biogeme*, Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

Bierlaire, M., Curchod, A., Danalet, A., Doyen, E., Faure, P., , A., Kaufmann, V., Tabaka, K. and Schuler, M. (2011). Projet de recherche sur la mobilité combinée, Rapport définitif de l'enquête de préférences révélées, *Technical report*, École Polytechnique Fédérale de Lausanne.

Bierlaire, M., Thémans, M. and Zufferey, N. (2009). A Heuristic for Nonlinear Global Optimization, *INFORMS Journal on Computing* **22**(1): 59–70.

Birkeland, M. E. and Jordal-Jorgensen, J. (2001). Energy efficiency of passenger cars, Cambridge, UK.

Boeri, M. (2011). *Advances in Stated Preference Methods: Discrete and Continuous Mixing Distributions in Logit Models for Representing Variance and Taste Heterogeneity*, PhD thesis, Queen's University Belfast.

Bollen, K. A. (1989). A New Incremental Fit Index for General Structural Equation Models, *Sociological Methods & Research* **17**(3): 303–316.

Brownstone, D. and Train, K. (1998). Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics* **89**(1–2): 109–129.

Cao, X., Mokhtarian, P. L. and Handy, S. L. (2006). Neighborhood design and vehicle type choice: Evidence from Northern California, *Transportation Research Part D: Transport and Environment* **11**(2): 133–145.

Cernicchiaro, G. and de Lapparent, M. (2015). A Dynamic Discrete/Continuous Choice Model for Forward-Looking Agents Owning One or More Vehicles, *Computational Economics* **46**(1): 15–34.

Chaniotakis, E. and Pel, A. J. (2015). Drivers' parking location choice under uncertain parking availability and search times: A stated preference experiment, *Transportation Research Part A: Policy and Practice* **82**: 228–239.

Choo, S. and Mokhtarian, P. L. (2004). What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice, *Transportation Research Part A: Policy and Practice* **38**(3): 201–222.

Chorus, C. G. (2015). Models of moral decision making: Literature review and research agenda for discrete choice analysis, *Journal of Choice Modelling* **16**: 69–85.

Chorus, C. G. and Kroesen, M. (2014). On the (im-) possibility of deriving transport policy implications from hybrid choice models, *Transport Policy* **36**: 217–222.

Coldren, G. M. and Koppelman, F. S. (2005). Modeling the competition among air-travel itinerary shares: GEV model development, *Transportation Research Part A: Policy and Practice* **39**(4): 345–365.

Comité des Constructeurs Français d'Automobiles (2016). SOeS; MEDDE; ASFA, Kantar Worldpanel; TNS; Setra; CPDP.

Crawford, G. S. (2000). The impact of the 1992 cable act on household demand and welfare, *SSRN Scholarly Paper ID 235290*, Social Science Research Network, Rochester, NY.

Daganzo, C. F. and Kusnic, M. (1993). Technical Note: Two Properties of the Nested Logit Model, *Transportation Science* **27**(4): 395–400.

Daziano, R. A. (2013). Conditional-logit Bayes estimators for consumer valuation of electric vehicle driving range, *Resource and Energy Economics* **35**(3): 429–450.

Daziano, R. A. and Achtnicht, M. (2014). Forecasting Adoption of Ultra-Low-Emission Vehicles Using Bayes Estimates of a Multinomial Probit Model and the GHK Simulator, *Transportation Science* **48**(4): 671–683.

de Jong, G., Fox, J., Daly, A., Pieters, M. and Smit, R. (2004). Comparison of car ownership models, *Transport Reviews* **24**(4): 379–408.

de Lapparent, M. and Cernicchiaro, G. (2012). How long to own and how much to use a car? A dynamic discrete choice model to explain holding duration and driven mileage, *Economic Modelling* **29**(5): 1737–1744.

de Palma, A. and Kilani, M. (2008). Regulation in the automobile industry, *International Journal of Industrial Organization* **26**(1): 150–167.

Dimitropoulos, A., Rietveld, P. and van Ommeren, J. N. (2013). Consumer valuation of changes in driving range: A meta-analysis, *Transportation Research Part A: Policy and Practice* **55**: 27–45.

Dupuit, J. (1844). On the measurement of the utility of public works, *International Economic Papers* **2**(1952): 83–110.

Dupuit, J. (1849). On tolls and transport charges, *Annales des ponts et chaussées*, Vol. 11, pp. 7–31.

European Comission (2011). Energy roadmap 2050. impact assessment and scenario analysis, `https://ec.europa.eu/energy/sites/ener/files/documents/roadmap2050_ia_20120430_en_0.pdf`.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*, Springer Netherlands, Dordrecht.

Forinash, C. V. and Koppelman, F. S. (1993). Application and Interpretation of Nested Logit Models of Intercity Mode Choice, *Transportation Research Record* (1413).

Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings, *Transportation Research Part B: Methodological* **40**(8): 688–707.

Frejinger, E. and Bierlaire, M. (2010). On Path Generation Algorithms for Route Choice Models, *in* S. Hess and A. Daly (eds), *Choice Modelling: The State-of-the-art and The State-of-practice*, Emerald Group Publishing Limited, pp. 307–315. DOI: 10.1108/9781849507738-013.

Glerum, A. (2014). *Static and dynamic mathematical models of behavior*, Thesis. École Polytechnique Fédérale de Lausanne.

Glerum, A., Atasoy, B. and Bierlaire, M. (2014). Using semi-open questions to integrate perceptions in choice models, *Journal of Choice Modelling* **10**: 11–33.

Glerum, A., Stankovikj, L., Thémans, M. and Bierlaire, M. (2014). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions, *Transportation Science* **48**(4): 483–499.

Golob, T. F. (2001). Joint models of attitudes and behavior in evaluation of the San Diego I-15 congestion pricing project, *Transportation Research Part A: Policy and Practice* **35**(6): 495–514.

Goolsbee, A. and Petrin, A. (2004). The consumer gains from direct broadcast satellites and the competition with cable TV, *Econometrica* **72**(2): 351–381.

Guan, W. (2003). From the help desk: Bootstrapped standard errors, *Stata Journal* **3**(1).

Guevara, C. A. (2010). *Endogeneity and Sampling of Alternatives in Spatial Choice Models*, Thesis, Massachusetts Institute of Technology. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Civil and Environmental Engineering, 2010.

Guevara, C. A. (2015). Critical assessment of five methods to correct for endogeneity in discrete-choice models, *Transportation Research Part A: Policy and Practice* **82**: 240–254.

Guevara, C. A. (2017). Mode-valued differences of in-vehicle travel time Savings, *Transportation* **44**(5): 977–997.

Guevara, C. A. and Ben-Akiva, M. (2012). Change of Scale and Forecasting with the Control-Function Method in Logit Models, *Transportation Science* **46**(3): 425–437.

Guevara, C. A. and Polanco, D. (2016). Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution, *Transportmetrica A: Transport Science* **12**(5): 458–478.

Hackbarth, A. and Madlener, R. (2013). Consumer preferences for alternative fuel vehicles: A discrete choice analysis, *Transportation Research Part D: Transport and Environment* **25**: 5–17.

Hackbarth, A. and Madlener, R. (2016). Willingness-to-pay for alternative fuel vehicle characteristics: A stated choice study for Germany, *Transportation Research Part A: Policy and Practice* **85**: 89–111.

Hasan, S., Ukkusuri, S., Gladwin, H. and Murray-Tuite, P. (2011). Behavioral Model to Understand Household-Level Hurricane Evacuation Decision Making, *Journal of Transportation Engineering* **137**(5): 341–348.

Hausman, J. A. (1978). Specification Tests in Econometrics, *Econometrica: Journal of the Econometric Society* **46**(6): 1251–1271.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system, *Working Paper 177*, National Bureau of Economic Research.

Heckman, J. J. and Robb Jr., R. (1985). Alternative methods for evaluating the impact of interventions: An overview, *Journal of Econometrics* **30**(1-2): 239–267.

Hess, S. and Axhausen, K. W. (2004). Checking our assumptions in value-of-travel-time modelling: Recovering taste distributions, *Technical report*, CTS Working Paper, Centre for Transport Studies, Imperial College London.

Hess, S., Fowler, M., Adler, T. and Bahreinian, A. (2012). A joint model for vehicle type and fuel type choice: evidence from a cross-nested logit study, *Transportation* **39**(3): 593–625.

Hoen, A. and Koetse, M. J. (2014). A choice experiment on alternative fuel vehicle preferences of private car owners in the Netherlands, *Transportation Research Part A: Policy and Practice* **61**: 199–215.

Hole, A. R. and Yoo, H. I. (2017). The use of heuristic optimization algorithms to facilitate maximum simulated likelihood estimation of random parameter logit models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Hunt, J. and Teply, S. (1993). A nested logit model of parking location choice, *Transportation Research Part B: Methodological* **27**(4): 253–265.

Imbens, G. (2014). Instrumental Variables: An Econometrician's Perspective, *Working Paper 19983*, National Bureau of Economic Research.

Institut national de la statistique et des études économiques (2016a). http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATTEF13629. Accessed on 20.11.2016.

Institut national de la statistique et des études économiques (2016b). http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATTEF05160. Accessed on 20.11.2016.

Institut national de la statistique et des études économiques (2016c). Prix moyens à la consommation en métropole - utilisation de véhicules, biens et services de loisirs, http://www.bdm.insee.fr/bdm2/affichageSeries?idbank=000442588&idbank=000849411\&bouton=OK&codeGroupe=169. Accessed on 21.11.2016.

Jara-Daz, S. R. and Videla, J. (1989). Detection of income effect in mode choice: Theory and application, *Transportation Research Part B: Methodological* **23**(6): 393–400.

Jensen, A. F., Cherchi, E., Mabit, S. L. and Ortzar, J. d. D. (2017). Predicting the Potential Market for Electric Vehicles, *Transportation Science* **51**(2): 427–440.

Jiang, R., Manchanda, P. and Rossi, P. E. (2009). Bayesian analysis of random coefficient logit models using aggregate data, *Journal of Econometrics* **149**(2): 136–148.

Johnston, M. A. and Paulsen, N. (2014). Rules of engagement: A discrete choice analysis of sponsorship decision making, *Journal of Marketing Management* **30**(7-8): 634–663.

Karaca-Mandic, P. and Train, K. (2003). Standard error correction in two-stage estimation with nestedsamples, *Econometrics Journal* **6**(2): 401–407.

Kim, J., Rasouli, S. and Timmermans, H. (2014). Expanding scope of hybrid choice models allowing for mixture of social influences and latent attitudes: Application to intended purchase of electric cars, *Transportation Research Part A: Policy and Practice* **69**: 71–85.

Knockaert, J. (2015). Estimating a latent class model: the issue of local optima, *Note*.

Koppelman, F. S. and Bhat, C. (2006). A self instructing course in mode choice modeling: multinomial and nested logit models.

Koppelman, F. S. and Hauser, J. R. (1978). Destination choice behavior for non-grocery-shopping trips, *Transportation Research Record* (673).

Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*, Wiley.

Lave, C. A. and Train, K. (1979). A disaggregate model of auto-type choice, *Transportation Research Part A: General* **13**(1): 1–9.

Li, W. (2014). *Modeling household residential choice using multiple imputation*, Thesis, Massachusetts Institute of Technology.

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T. A., Carson, R. T., Deshazo, J. R., Fiebig, D. and Greene, W. (2005). Recent progress on endogeneity in choice modeling, *Marketing Letters* **16**(3-4): 255–265.

Lurkin, V. (2016). *Modeling in Air Transportation: Cargo Loading and Itinerary Choice*, PhD thesis, Universitè de Liège, Liège, Belgique.

Mai, T., Frejinger, E., Fosgerau, M. and Bastin, F. (2015). A Dynamic Programming Approach for Quickly Estimating Large MEV Models, *Technical report*.

Massiani, J. (2014). Stated preference surveys for electric and alternative fuel vehicles: are we doing the right thing?, *Transportation Letters* **6**(3): 152–160.

McCarthy, P. S. and Tay, R. S. (1998). New Vehicle Consumption and Fuel Efficiency: A Nested Logit Approach, *Transportation Research Part E: Logistics and Transportation Review* **34**(1): 39–51.

McFadden, D. (1986). The Choice Theory Approach to Market Research, *Marketing Science* **5**(4): 275–297.

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

Ministère de l'environnement, de l'énergie et de la mer (2016). Bonus-malus : définitions et barèmes pour 2016, `http://www.developpement-durable.gouv.fr/Bonus-Malus-definitions-et-baremes.html`. Accessed on 21.11.2016.

Mohammadian, A. and Miller, E. (2002). Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, *Transportation Research Record: Journal of the Transportation Research Board* **1807**: 92–100.

Mohammadian, A. and Miller, E. (2003). Empirical Investigation of Household Vehicle Type Choice Decisions, *Transportation Research Record: Journal of the Transportation Research Board* **1854**: 99–106.

N. Flynn, T., J. Peters, T. and Coast, J. (2013). Quantifying response shift or adaptation effects in quality of life by synthesising best-worst scaling and discrete choice data, *Journal of Choice Modelling* **6**: 34–43.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry, *Econometrica* **69**(2): 307–342.

Newey, W. K. (1985). Semiparametric estimation of limited dependent variable models with endogenous explanatory variables, *Annales de l'inséé* (59/60): 219–237.

Newey, W. K. (1987). Efficient estimation of limited dependent variable models with endogenous explanatory variables, *Journal of Econometrics* **36**(3): 231–250.

Office fédéral de la statistique, O. f. d. d. t. (2012). *La mobilité en Suisse: Résultats du microrecensement mobilité et transports 2010*, Office fédéral de la statistique, Neuchâtel.

Olson, J. M. and Zanna, M. P. (1993). Attitudes and Attitude Change, *Annual Review of Psychology* **44**(1): 117–154.

Outwater, M., Castleberry, S., Shiftan, Y., Ben-Akiva, M., Shuang Zhou, Y. and Kuppam, A. (2003). Attitudinal Market Segmentation Approach to Mode Choice and Ridership Forecasting: Structural Equation Modeling, *Transportation Research Record: Journal of the Transportation Research Board* **1854**: 32–42.

Ozdemir, S., Johnson, F. R. and Whittington, D. (2016). Ideology, public goods and welfare valuation: An experiment on allocating government budgets, *Journal of Choice Modelling* **20**: 61–72.

Pacheco, M., Azadeh, S. S., Bierlaire, M. and Gendron, B. (2017). Integrating advanced discrete choice models in mixed integer linear optimization, *Technical Report TRANSP-OR 170714*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Page, M., Whelan, G. and Daly, A. (2000). Modelling the factors which influence new car purchasing.

Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas, *Marketing Science* **31**(4): 567–586.

Pels, E., Nijkamp, P. and Rietveld, P. (2003). Access to and competition between airports: a case study for the San Francisco Bay area, *Transportation Research Part A: Policy and Practice* **37**(1): 71–83.

Petrin, A. and Train, K. (2003). Omitted Product Attributes in Discrete Choice Models, *Working Paper 9452*, National Bureau of Economic Research.

Petrin, A. and Train, K. (2009). A control function approach to endogeneity in consumer choice models, *Journal of Marketing Research* **47**(1): 3–13.

Potoglou, D. (2008). Vehicle-type choice and neighbourhood characteristics: An empirical study of Hamilton, Canada, *Transportation Research Part D: Transport and Environment* **13**(3): 177–186.

Proussaloglou, K. E. and Koppelman, F. S. (1989). Use of travelers' attitudes in rail service design, *Transportation Research Record* (1221).

Qin, D. (2015). Resurgence of the Endogeneity-Backed Instrumental Variable Methods, *Economics: The Open-Access, Open-Assessment E-Journal* .

Rasouli, S. and Timmermans, H. (2016). Influence of Social Networks on Latent Choice of Electric Cars: A Mixed Logit Specification Using Experimental Design Data, *Networks and Spatial Economics* **16**(1): 99–130.

Rodrigues, L. C., van den Bergh, J. C. J. M., Loureiro, M. L., Nunes, P. A. L. D. and Rossi, S. (2016). The Cost of Mediterranean Sea Warming and Acidification: A Choice Experiment Among Scuba Divers at Medes Islands, Spain, *Environmental and Resource Economics* **63**(2): 289–311.

Sadri, A. M., Ukkusuri, S. V. and Gladwin, H. (2017). Modeling joint evacuation decisions in social networks: The case of Hurricane Sandy, *Journal of Choice Modelling* .

Schafer, J. L. (2000). *Analysis of incomplete multivariate data*, number 72 in *Monographs on statistics and applied probability*, 1. ed., 1. crc press reprint edn, Chapman & Hall/CRC, Boca Raton. OCLC: 249266966.

Schlereth, C. (2014). Pricing plans for a financial advisory service, *European Journal of Marketing* **48**(3/4): 595–616.

Sud Ouest (2015). Prix du gazole et de l'essence : ce qui va changer pour les automobilistes, http://www.sudouest.fr/2015/10/14/carburant-le-gouvernement-annonce-l-augmentation-de-la-taxe-sur-le-gazole-des-2016-2154678-4755.php. Accessed on 21.11.2016.

Train, K. (1980). The potential market for non-gasoline-powered automobiles, *Transportation Research Part A: General* **14**(5-6): 405–414.

Train, K. (1986). *Qualitative choice analysis: theory, econometrics, and an application to automobile demand*, number 10 in *MIT Press series in transportation studies*, MIT Press, Cambridge, Mass.

Train, K. (2003). *Discrete Choice Methods with Simulation*, Cambridge University Press.

Train, K. (2009). *Discrete choice methods with simulation*, 2nd ed edn, Cambridge University Press, Cambridge ; New York. OCLC: ocn349248337.

Train, K. E. and Winston, C. (2007a). Vehicle choice behavior and the declining market share of US automakers, *International Economic Review* **48**(4): 1469–1496.

Train, K. E. and Winston, C. (2007b). Vehicle choice behavior and the declining market share of U.S. automakers, *International Economic Review* **48**(4): 1469–1496.

U.S. Department of Energy (2016). Energy efficiency & renewable energy, `http://www.fueleconomy.gov/feg/`. Accessed on 21.11.2016.

Villas-Boas, J. M. and Winer, R. S. (1999). Endogeneity in brand choice models, *Management Science* **45**(10): 1324–1338.

Vovsha, P. and Bekhor, S. (1998). Link-Nested Logit Model of Route Choice: Overcoming Route Overlapping Problem, *Transportation Research Record: Journal of the Transportation Research Board* **1645**: 133–142.

Vredin Johansson, M., Heldt, T. and Johansson, P. (2006). The effects of attitudes and personality traits on mode choice, *Transportation Research Part A: Policy and Practice* **40**(6): 507–525.

Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model, *Mathematical Social Sciences* **43**(3): 303–343.

Walker, J. L. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables*, PhD thesis, Massachusetts Institute of Technology.

Walker, J. L., Ehlers, E., Banerjee, I. and Dugundji, E. R. (2011). Correcting for endogeneity in behavioral choice models with social influence variables, *Transportation Research Part A: Policy and Practice* **45**(4): 362–374.

Wang, S., Fan, J., Zhao, D., Yang, S. and Fu, Y. (2016). Predicting consumers' intention to adopt hybrid electric vehicles: using an extended version of the theory of planned behavior model, *Transportation* **43**(1): 123–143.

Webb, J., Briggs, M. and Wilson, C. (2017). Breaking automotive modal lock-in: a choice modelling study of Jakarta commuters, *Environmental Economics and Policy Studies* .

Wong, T., Brownstone, D. and Bunch, D. S. (2017). Aggregation Biases in Discrete Choice Models, *Technical report*.

Wood, W. (2000). Attitude Change: Persuasion and Social Influence, *Annual Review of Psychology* **51**(1): 539–570.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Wu, G., Yamamoto, T. and Kitamura, R. (1999). Vehicle Ownership Model That Incorporates the Causal Structure Underlying Attitudes Toward Vehicle Ownership, *Transportation Research Record: Journal of the Transportation Research Board* **1676**: 61–67.

Yang, C.-W. and Wang, H.-C. (2017). A comparison of flight routes in a dual-airport region using overlapping error components and a cross-nested structure in GEV models, *Transportation Research Part A: Policy and Practice* **95**: 85–95.

Yang, S., Chen, Y. and Allenby, G. M. (2003). Bayesian analysis of simultaneous demand and supply, *Quantitative Marketing and Economics* **1**(3): 251–275.

Yao, S., Wang, W. and Chen, Y. (2016). TV Channel Search and Commercial Breaks, *Journal of Marketing Research* p. jmr.15.0121.

Zolfaghari, A. (2013). Methodological and empirical challenges in modelling residential location choices.

Anna Fernández Antolín

Transport and Mobility Laboratory – TRANSP-OR
School of Architecture, Civil and Environmental Engineering, Station 18
Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
☎ (+41) 21 69 32435 • ✉ anna.fernandez.antolin@gmail.com
website: http://transp-or.epfl.ch/personnal.php?Person=FERNANDEZ

Education

École Polytechnique Fédérale de Lausanne — **Lausanne**
Doctoral degree (PhD) — *October 2013–present*
Dealing with Correlations in Discrete Choice Models.

Universitat Politècnica de Catalunya — **Barcelona**
MSc in Mathematics — *2007-2013*

École Polytechnique Fédérale de Lausanne — **Lausanne**
Erasmus exchange — *2012-2013*
One semester exchange programme

Oak House British School — **Barcelona**
High School — *2005-2007*
Final mark: 9/10

Professional Experience

Novartis — **Basel**
Modeling and simulation intern — *2013*
Modeling of drug release from liposomes

École Polytechnique Fédérale de Lausanne — **Lausanne**
Transp-or Laboratory intern — *2013*
A dynamic vehicle choice forecasting framework

Teaching mathematics extra lessons — **Barcelona**
Private teacher for secondary and high school students — *2008-2012*

Club de Natació Barcelona — **Barcelona**
Swimming instructor — *Summers of 2006,2007*

Languages

English: Full professional proficiency — *Certificate of Proficiency in English, Cambridge*
French: Professional working proficiency
German: Elementary proficiency — *Taking a B1 course in the Language Center of EPFL*
Spanish: Native — *Mother Tongue*
Catalan: Native — *Mother Tongue*

Computer skills

Maths: MatLab
Statistics: R
Languages: C++, java
Other: LaTeX, pythonbiogeme

Research projects

Nissan **Lausanne**
Electric vehicle adoption dynamics: exploring market potentials *2014–2017*
This project proposes innovative methods to identify the determinants of acceptance of alternative vehicles and their impact on everyday mobility.

Teaching

Mathematical modeling of behavior
EPFL, Master course *2013–2017*

Decision-aid methodologies in transportation
EPFL, Master course *2014-2016*

Discrete Choice Analysis: Predicting Demand and Market Shares
EPFL, Postgraduate one week program *2013–2017*

Student project supervision

Master theses

1. *Discrete choice model for the automobile market: modeling willingness-to-pay for marginal and non marginal changes in car attributes (2016).* Anna-Katharina Clodong (EPFL).
2. *Disaggregate modeling of vehicle-miles traveled: what about users of alternative-fuel vehicles? (2015).* Takao Dantsuji (EPFL & Tokyo Institute of Technology).
3. *Modeling consideration and willingness-to-pay for electric and plug-in hybrid electric vehicles in car renewal (2015).* Maurin Baillif (EPFL).
4. *Mode Choice Analysis Using Smartphone Data (2014).* Michael Friederich (EPFL).

Semester projects

1. *Hedonic pricing of car attributes: a comparison across European countries (2015).* Anna-Katharina Clodong (EPFL).
2. *Accounting for attitudes in modeling demand for electric vehicles (2014).* Maurin Baillif (EPFL).
3. *Investigating the role of attitudes in the purchase of new cars (2016).* Nicola Ortelli (EPFL).
4. *Modeling purchases of new cars for 2015: a comparison between countries (2017).* Martí Montesinos (EPFL).
5. *Integrating demand and supply in the context of airlines (2017).* Gabriel Curis and Thibaut Richard (EPFL).
6. *Activity based modeling (2017).* Nicola Ortelli (EPFL).

Publications

Papers in international journals

1. Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data, *Journal of Choice Modelling* **20**:1-15. doi:10.1016/j.jocm.2016.09.001
2. Fernandez-Antolin, A., de Lapparent, M., and Bierlaire, M. (2017). Modeling purchases of new cars: an analysis of the 2014 French market, *Theory and Decision*. doi:10.1007/s11238-017-9631-y

Papers in conference proceedings

1. Fernández-Antolín, A., Lurkin, V., de Lapparent, M., and Bierlaire, M. (2017). Discrete-continuous maximum likelihood for the estimation of nested logit models. Proceedings of the 16th Swiss Transport Research Conference (STRC) 17-19 May, 2017.
2. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M. (2016). Uncovering substitution patterns in new car sales using a cross nested logit model. Proceedings of the 16th Swiss Transport Research Conference (STRC) 18-20 May, 2016.
3. Baillif, M., de Lapparent, M., Fernández-Antolín, A., and Bierlaire, M. (2015). Modeling consideration for electric vehicles and plug-in electric vehicles in car renewal. Proceedings of the 15th Swiss Transportation Research Conference (STRC) April 15-17, 2015.
4. Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M. (2015). Correcting for endogeneity using the EMIS method: a case study with revealed preference data. Proceedings of the 15th Swiss Transport Research Conference (STRC) 15-17 April, 2015.
5. Fernández-Antolín, A., Stathopoulos, A., and Bierlaire, M. (2014). Exploratory Analysis of Endogeneity in Discrete Choice Models. Proceedings of the 14th Swiss Transport Research Conference (STRC) 14-16 May, 2014.

Technical reports

1. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M. (2016). Modeling purchases of new cars: an analysis of the 2014 French market. Technical report TRANSP-OR 161130.
2. Fernández-Antolín, A., Guevara, C.A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity using the multiple indicator solution. Technical report TRANSP-OR 160405.

Presentations at international conferences & invited seminars

1. Fernández-Antolín, A., Lurkin, V., and Bierlaire, M., Discrete-continuous maximum likelihood estimation. 2017 INFORMS Annual Meeting, October 23, 2017, Houston, USA
2. Fernández-Antolín, A., Lurkin, V., and Bierlaire, M., Discrete-continuous maximum likelihood. 6th Symposium of the European Association for Research in Transportation, Faculty of Civil and Environmental Engineering, Technion, September 14, 2017, Haifa, Israel
3. Fernández-Antolín, A., Lurkin, V., Pacheco, M., and Bierlaire, M., Discrete-continuous maximum likelihood. Workshop on Discrete Choice Models 2017, EPFL, June 23, 2017, Lausanne, Switzerland
4. Fernández-Antolín, A., Lurkin, V., de Lapparent, M., and Bierlaire, M., Discrete-continuous maximum likelihood for the estimation of nested logit models. 16th Swiss Transport Research Conference (STRC), May 18, 2017, Monte Verita, Ascona, Switzerland
5. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M., Modeling demand for new cars in France. hEART 2016, 5th Symposium of the European Association for Research in Transportation, Delft University of Technology, September 14, 2016, Delft, Netherlands
6. Fernández-Antolín, A., How do we make a choice?. My Thesis in 180 Seconds, Ecole Polytechnique Fédérale de Lausanne, September 06, 2016, Lausanne, Switzerland Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M., Flexible substitution patterns in new car sales. TRISTAN IX, June 13, 2016, Oranjestad, Aruba
7. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M., Uncovering substitution patterns in new car sales using a cross nested logit model. 16th Swiss Transport Research Conference (STRC), May 19, 2016, Monte Verita, Ascona, Switzerland
8. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M., How do I choose a new car? A discrete-choice approach.. Invited lecture, Ecole des Ponts - ParisTech, May 12, 2016, Paris, France
9. Fernández-Antolín, A., de Lapparent, M., and Bierlaire, M., I would like a new car, which one do I choose?. Workshop on Discrete Choice Models 2016, April 22, 2016, Lausanne, Switzerland
10. Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M., Capturing non-linearities between

observed and unobserved variables: how to model the "car loving" attitude while correcting for endogeneity. 4th Symposium of the European Association for Research in Transportation, Technical University of Denmark, September 09, 2015, Copenhagen, Denmark

11. Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M., Modeling the unobserved: an application of the MIS method to RP Swiss data. 14th International Conference on Travel Behaviour Research, July 19, 2015, Beaumont Estate, Windsor
12. Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M., Correcting for endogeneity using the MIS method: a case study using mode choice RP data in Switzerland. Workshop on Discrete Choice Models 2015, May 28, 2015, Lausanne, Switzerland
13. Fernández-Antolín, A., Guevara, C.A., and Bierlaire, M., Correcting for endogeneity using the EMIS method: a case study with revealed preference data. 15th Swiss Transport Research Conference (STRC), April 16, 2015, Monte Verita, Ascona, Switzerland
14. Guevara, C.A., Fernández-Antolín, A., Glerum, A., and Bierlaire, M., A correction for endogeneity in choice models with psychological constructs. 3rd Symposium of the European Association for Research in Transportation (hEART) 2014, Institute for Transport Studies, University of Leeds, September 11, 2014, Leeds, United Kingdom
15. Fernández-Antolín, A., Bierlaire, M., Fosgerau, M., and McFadden, D., Choice Probability Generating Functions. 9th Workshop on Discrete Choice Models, June 20, 2014, Lausanne, Switzerland
16. Fernández-Antolín, A., Stathopoulos, A., and Bierlaire, M., Exploratory Analysis of Endogeneity in Discrete Choice Models . 14th Swiss Transportation Research Conference, May 15, 2014, Ascona, Switzerland

Article reviewing

European Transport Research Review

Journal of Urban Planning and Developement

Transportation

Transportation Research Part A: Policy and Practice

Miscellaneous

Lausanne Natation — **Lausanne**
Member of the waterpolo committee — *2016–Present*

Swiss swimming — **Switzerland**
National waterpolo referee — *2015–Present*

Cercle des Nageurs de Nyon — **Nyon**
Waterpolo player — *2014–Present*
Women Swiss League

Board of European Students of Technology — **Madrid**
Winner of the national Engineering Competition — *2012*

PACCS — **Barcelona**
Leisure Monitor — *2008-2012*
In charge of a group of teenagers along with a team of monitors

Club de Natació Sant Feliu — **Barcelona**
Waterpolo — *2004-2008*
Highest Spanish league

Other information

○ Swiss residence permit type B, driving licence types A and B.