



AMERICAN UNIVERSITY OF BEIRUT

INTEGRATION OF MACHINE LEARNING AND DISCRETE  
CHOICE MODELS TO BETTER PREDICT AND DESCRIBE  
DECISION MAKERS' CHOICES WITH APPLICATIONS TO  
TRAVEL DECISIONS

by  
GEORGES MAROUN SFEIR

A dissertation  
submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy  
to the Department of Civil and Environmental Engineering  
of the Maroun Semaan Faculty of Engineering and Architecture  
at the American University of Beirut

Beirut, Lebanon  
April 2021

AMERICAN UNIVERSITY OF BEIRUT

INTEGRATION OF MACHINE LEARNING AND DISCRETE  
CHOICE MODELS TO BETTER PREDICT AND DESCRIBE  
DECISION MAKERS' CHOICES WITH APPLICATIONS TO  
TRAVEL DECISIONS

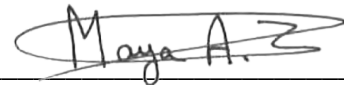
by  
GEORGES MAROUN SFEIR

Approved by:

---

Dr. Salah Sadek, Professor  
Department of Civil and Environmental Engineering

Chair of Committee



---

Dr. Maya Abou-Zeid, Associate Professor  
Department of Civil and Environmental Engineering

Advisor

---

Dr. Isam Kaysi, Adjunct Professor  
Department of Civil and Environmental Engineering

Co-Advisor

---

Dr. Francisco Camara Pereira, Professor  
Department of Technology, Management and Economics  
Technical University of Denmark

Co-Advisor

---

Dr. Filipe Rodrigues, Associate Professor  
Department of Technology, Management and Economics  
Technical University of Denmark

Member of Committee

---

Dr. Mariette Awad, Associate Professor  
Department of Electrical and Computer Engineering

Member of Committee

---

Dr. Ali Chalak, Associate Professor  
Department of Agriculture

Member of Committee

---

Dr. Bilal Farooq, Associate Professor  
Department of Civil Engineering  
Ryerson University

Member of Committee

Date of dissertation defense: April 22, 2021



## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Professor Maya Abou-Zeid, my research advisor, for her guidance, patience, professionalism, and constant encouragement throughout my PhD journey. I thank her for always being available to answer my questions, for her drive for perfection, for pushing me further in research, and for broadening my research horizons. I am lucky to be her student.

I extend my gratitude to my co-advisors. Thanks to Professor Isam Kaysi for his valuable insights and feedback, for providing me with great professional experiences, for his time to answer my questions, for all the discussions we had, and for his sense of humor. Thanks to Professor Francisco Camara Pereira for giving me the opportunity to join the Machine Learning for Smart Mobility (MLSM) group at the Technical University of Denmark (DTU). I also thank him for his feedback and guidance. I am honored to be your student.

I also extend my gratitude to Professor Filipe Rodrigues for his patience, feedback, and suggestions to further improve the models and algorithms I used in this dissertation. Thank you for your support and for answering my endless questions.

I would also like to thank Professor Salah Sadek for leading my doctoral committee and reviewing my work. I am also thankful to all the members of my doctoral committee: Professor Mariette Awad for reviewing this dissertation and for her Pattern Recognition course that introduced me to the world of machine learning and inspired this dissertation, Professor Bilal Farooq for reviewing this dissertation and whose work was an inspiration to me, and Professor Ali Chalak for his interest in my research and reviewing my work.

Thanks to the Dean's office of the Maroun Semaan Faculty of Engineering and Architecture (MSFEA) at AUB for funding my research. Thanks to the Civil and Environmental Engineering Department and all the administrative and laboratory staff, especially Mrs. Zakeya Deeb, Mr. Helmi Al-Khatib, and Ms. Dima Al-Hassanieh. Thanks also to the MSFEA IT unit: Mr. Sami El Ghossaini, Mr. Ali Kaafarani, Mr. Tarek Bou Hamdam, and Mr. Youssef Yehia.

Many thanks to my AUB labmates: Najib Zgheib, Lara Otary, Hovsep Apkarian, Rana Tarabay, Marianne Jreige, and Maher Said, with whom I have shared many memorable moments and long days at AUB. Thank you for being supportive.

I am also grateful to all the members of the MLSM group at DTU: Professor Carlos Lima Azevedo, Professor Stanislav Borysov, Daniele Gammelli, Valentino Servizi, Sergio Garrido, Lampros Yfantis, Ioulia Markou, Mayara Monteiro, Bojan Kostic, Niklas Petersen. Thanks to Daniele Daler, Marcus Myhrmann, Morten Eltved, Luca Furlanetto, Andrea Papu, Daniele Ciardi, Giovanni Albano and all my friends at DTU.

Thanks to my friends who have supported me throughout this journey. Wissam, thank you for always being there and for your endless support. Anthony, thank you for all the brainwashing sessions that encouraged me to pursue a PhD at AUB. Jimmy, thank you for sharing our PhD journeys and struggles together. Daniele Gammelli, thank you for all the help during my visit to DTU and for all the pizzas. Thanks also to Jospheh A., Joseph M., Elissa, Daniel, Tera, Rafy, Zeina, Emanuel, Grace-Mary, Dr. Elizabeth & my uncle Hares, Dr. Gilbert & Hanadi, Dr. Joe & Sandy, Grace, Ghinwa and every member of my big family for all your support throughout the years.

Finally, I cannot thank enough my mom and dad for their support and love. I also cannot thank enough my amazing brothers, Elie, Pierre & Saidy, and Joe for their endless support and patience.

# ABSTRACT OF THE DISSERTATION OF

Georges Maroun Sfeir

for

Doctor of Philosophy

Major: Civil and Environmental Engineering

Title: Integration of Machine Learning and Discrete Choice Models to Better Predict and Describe Decision Makers' Choices with Applications to Travel Decisions

This dissertation develops methods that combine the advantages of discrete choice models and machine learning methods into interpretable econometric models. The aim is to enhance the predictive power of discrete choice models and their flexibility in representing unobserved heterogeneity without weakening their behavioral and economic interpretability. Specifically, this dissertation focuses on bringing machine learning into the Latent Class Choice Models (LCCMs), which are widely used in the discrete choice modeling community to model the unobserved behavioral heterogeneity of a population through discrete segments (or latent classes). LCCM consists of two sub-components, a class membership model that formulates the probability of an individual belonging to a specific segment/class and a class-specific choice model that estimates the choice probabilities.

The dissertation develops two new Latent Class Choice Models with a flexible class membership component. In each of the two proposed models, the latent classes are defined using a different machine learning clustering technique as opposed to the random utility specification of the LCCM. The first proposed model is titled Gaussian-Bernoulli Mixture – Latent Class Choice Model (GBM-LCCM) while the second proposed model is called Gaussian Process – Latent Class Choice Model (GP-LCCM).

The GBM-LCCM formulates the latent classes using model-based mixture models as an alternative approach to the traditional random utility specification with the aim of comparing the two approaches on various measures including prediction accuracy and representation of heterogeneity in the choice process. Mixture models are parametric model-based clustering techniques that have been widely used in areas such as machine learning, data mining and pattern recognition for clustering and classification problems. An Expectation-Maximization (EM) algorithm is derived for the estimation of the proposed model. Using two different case studies on travel mode choice behavior, the proposed model is compared to its traditional discrete choice model counterpart, the LCCM, on the basis of parameter estimate signs, values of time, statistical goodness-of-fit measures, and cross-validation tests. Results show that mixture models improve the overall performance of LCCMs by providing better out-of-sample prediction accuracy by around 3% in addition to better and more flexible representation of heterogeneity



and more reasonable parameter estimate signs without weakening the behavioral and economic interpretability of the choice models.

The second model, the GP-LCCM, formulates the latent classes using Gaussian Processes (GPs), a nonparametric class of probabilistic machine learning. Gaussian Processes are kernel-based algorithms that incorporate expert knowledge by assuming priors over latent functions rather than priors over parameters, which makes them more flexible in addressing nonlinear problems. By integrating a Gaussian Process within the LCCM structure, we aim at improving discrete representations of unobserved heterogeneity. The proposed model would assign individuals probabilistically to behaviorally homogeneous clusters (latent classes) using GPs and simultaneously estimate class-specific choice models by relying on random utility models. Furthermore, we derive and implement an Expectation-Maximization algorithm to jointly estimate/infer the hyper-parameters of the GP kernel function and the class-specific choice parameters by relying on a Laplace approximation and gradient-based numerical optimization methods, respectively. The model is tested on three different mode choice applications and compared against the traditional LCCM and the proposed GBM-LCCM. Results show that the GP-LCCM allows for a more complex and flexible representation of heterogeneity and improves both in-sample fit and out-of-sample predictive power by up to 7.6% and 8.8%, respectively. Moreover, behavioral and economic interpretability is maintained at the class-specific choice model level while local interpretation of the latent classes can still be achieved, although the nonparametric characteristic of GPs lessens the transparency of the class membership component.

The two proposed models are also compared against the LCCM in terms of their forecasting capabilities. Results show that both the GBM-LCCM and GP-LCCM are capable of providing meaningful forecasts that are similar to the forecasts of the traditional LCCM, to some extent. A demand sensitivity analysis with respect to the cost of some travel mode alternatives is also conducted and similar order of changes are attained between the results of the proposed models and LCCM in terms of in-sample fit, out-of-sample prediction accuracy, and aggregate forecasts. The sensitivity analysis also highlights the advantage of the proposed models in identifying a higher number of classes than the LCCM by providing a more in-depth understanding of the behavioral heterogeneity within a population and the behavioral responses of the different classes to new policies.

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS .....	1
ABSTRACT .....	3
ILLUSTRATIONS .....	9
TABLES .....	11
INTRODUCTION .....	12
1.1. Motivation.....	12
1.1.1. Discrete Choice Models.....	13
1.1.2. Machine Learning in Choice Modeling .....	15
1.2. Research Objectives.....	16
1.3. Research Contributions.....	18
1.3.1. Gaussian-Bernoulli Mixture – Latent Class Choice Model (GBM-LCCM)	18
1.3.2. Gaussian Process – Latent Class Choice Model (GP-LCCM) .....	18
1.3.3. Estimation and Application .....	19
1.4. Outline .....	20
LITERATURE REVIEW .....	22
2.1. Discrete Choice Model .....	22
2.1.1. MNL Formulation.....	23
2.1.2. Taste Heterogeneity .....	26
2.2. Machine Learning in Choice Modeling.....	32
2.2.1. General Description .....	33

2.2.2. Comparative Studies .....	34
2.2.3. Differences .....	39
2.3. Combining the Two Fields .....	44
2.4. This Dissertation in Context .....	45
<b>GAUSSIAN-BERNOULLI MIXTURE LATENT CLASS CHOICE MODEL .....</b>	<b>50</b>
3.1. Latent Class Choice Model.....	51
3.2. Gaussian-Bernoulli Mixture Latent Class Choice Model.....	53
3.2.1. EM Algorithm.....	58
3.2.2. Final Likelihood.....	61
<b>GBM-LCCM APPLICATIONS.....</b>	<b>63</b>
4.1. Data.....	63
4.1.1. London Dataset .....	63
4.1.2. AUB Dataset .....	64
4.2. Implementation .....	65
4.3. Applications .....	67
4.3.1. London Case Study .....	67
4.3.2. AUB Case Study .....	77
4.4. Conclusion .....	86
<b>GAUSSIAN PROCESS LATENT CLASS CHOICE MODEL..</b>	<b>88</b>
5.1. Gaussian Process.....	89
5.2. Kernels .....	91
5.2.1. Squared Exponential Kernel (SE) / Radial Basis Function (RBF) .....	91

5.2.2. Matérn Kernel .....	92
5.3. Gaussian Process – Latent Class Choice Model .....	93
5.3.1. Proposed Model .....	93
5.3.2. EM Algorithm .....	95
<b>GP-LCCM APPLICATIONS .....</b>	<b>99</b>
6.1. Data .....	99
6.1.1 Swissmetro Dataset .....	99
6.2. Implementation .....	100
6.3. Applications .....	101
6.3.1. AUB Case Study .....	101
6.3.2. London Case Study .....	110
6.3.3. Swissmetro Case Study .....	116
6.4. Conclusion .....	119
<b>POLICY ANALYSIS .....</b>	<b>122</b>
7.1. Discrete Choice Models and Machine Learning for Policy Analysis .....	122
7.2. Application .....	125
7.3. Results .....	126
7.3.1. Base Case Scenario .....	126
7.3.2. Sensitivity Analysis .....	132
7.4. Calibration of the Constants .....	143
7.5. Conclusion .....	144
<b>CONCLUSION .....</b>	<b>146</b>

8.1. Summary .....	146
8.2. Limitations and Future Directions .....	148
8.3. Conclusion .....	151
<b>APPENDIX A: LONDON DATASET .....</b>	<b>153</b>
A.1. London Dataset – First Specification.....	153
A.1.1. Latent Class Choice Models .....	153
A.1.2. Gaussian-Bernoulli Mixture - Latent Class Choice Models .....	155
A.2. London Dataset – Second Specification .....	162
A.2.1. Latent Class Choice Models .....	162
A.2.2. Gaussian-Bernoulli Mixture - Latent Class Choice Models .....	164
A.3. London Dataset – Third Specification .....	171
A.3.1. Gaussian-Bernoulli Mixture - Latent Class Choice Models .....	171
<b>APPENDIX B: AUB DATASET .....</b>	<b>175</b>
B.1. Full Covariance .....	175
B.2. Tied Covariance .....	176
B.3. Diagonal Covariance .....	177
B.4. Spherical Covariance .....	178
<b>APPENDIX C: SWISSMETRO DATASET .....</b>	<b>180</b>
C.1. Latent Class Choice Model .....	180
C.2. Gaussian Process - Latent Class Choice Model .....	182
<b>REFERENCES .....</b>	<b>183</b>

## ILLUSTRATIONS

Figure	Page
1. Latent Class Choice Model Framework (adapted from Walker and Li (2007)).	29
2. Gaussian-Bernoulli Mixture - Latent Class Choice Model (GBM-LCCM) .....	54
3. Graphical Representation of the proposed Gaussian-Bernoulli Mixture Latent Class Choice Model (GBM-LCCM) for a set of $N$ decision-makers .....	54
4. Hypothetical scenario and choice question example from the survey .....	65
5. Graphical representation of the proposed Gaussian Process – Latent Class Choice Model (GP-LCCM) for a set of $N$ decision-makers and $K$ clusters/latent classes .....	93
6. Explaining three individual class predictions of the GP-LCCM with two classes ( $K = 2$ ) using LIME .....	110
7. Explaining two individual class predictions of the GP-LCCM with two classes ( $K = 2$ ) using LIME .....	116
8. Demand for a product or service as a function of its price (Aboutaleb et al. (2021)) .....	124
9. Expected weekly mode share of individuals under the base case scenario .....	127
10. Expected weekly trips per mode under the base case scenario.....	128
11. Expected weekly mode share of individuals per class under the base case scenario of LCCM ( $K = 2$ ).....	129
12. Expected weekly trips per mode and class under the base case scenario of LCCM ( $K = 2$ ) .....	129
13. Expected weekly mode share of individuals per class under the base case scenario of GP-LCCM ( $K = 3$ ) .....	131
14. Expected weekly trips per mode and class under the base case scenario of GP-LCCM ( $K = 3$ ) .....	131
15. Percent difference of weekly shared-taxi trips as forecasted by LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shared-taxi .....	133
16. Percent difference of weekly shuttle trips as forecasted by LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shared-taxi .....	133

17. Percent difference of weekly shared-taxi trips as forecasted by LCCM (K = 2) and GP-LCCM (K = 3) with respect to changes in the one-way fare of shuttle .....	134
18. Percent difference of weekly shuttle trips as forecasted by LCCM (K = 2) and GP-LCCM (K = 3) with respect to changes in the one-way fare of shuttle .....	135
19. Expected weekly mode(s) share of individuals per class under different one-way shared-taxi fares - LCCM (K = 2) .....	137
20. Expected weekly trips per mode and class under different one-way shared-taxi fares – LCCM (K = 2).....	138
21. Expected weekly mode(s) share of individuals per class under different one-way shared-taxi fares – GP-LCCM (K = 3) .....	141
22. Expected weekly trips per mode and class under different one-way shared-taxi fares – GP-LCCM (K = 3) .....	142

## TABLES

Table	Page
1. Equivalent terminology between Machine Learning and Econometric Models (adapted from Sarle (1994); Karlaftis and Vlahogianni (2011); Hillel et al., (2020)) .....	40
2. EM Initialization .....	67
3. First trial.....	74
4. Mean matrix of the class membership model (GBM) – Tied Covariance – $K = 4$ .....	75
5. Parameter estimates of the class-specific choice models – Tied Covariance – $K = 4$ .....	75
6. Second trial .....	76
7. Third trial .....	76
8. Explanatory variables used in the models.....	79
9. Summary results of LCCM and GM-LCCM.....	83
10. LCCM – $K = 2$ .....	84
11. GM-LCCM – $K = 2$ .....	84
12. GM-LCCM - $K = 3$ .....	85
13. Summary results of the AUB application .....	103
14. Class-specific choice models and VOT ( $K = 2$ ) .....	105
15. Class-specific choice models ( $K = 3$ ) .....	106
16. Class membership estimates of LCCM and GM-LCCM ( $K = 2$ ).....	107
17. Summary results of the London application .....	114
18. Class-specific choice estimates of GBM-LCCM and GP-LCCM ( $K = 2$ ) .....	115
19. Mean matrix of the class membership model of GBM-LCCM ( $K = 2$ ) .....	115
20. Variables used to define the latent classes.....	117
21. Summary results of LCCM.....	118
22. Summary results of GP-LCCM .....	119
23. Attributes of the shared-taxi and shuttle options (adapted from Sfeir et al. (2020)) .....	126



# CHAPTER 1

## INTRODUCTION

This dissertation examines the possibility of combining Machine Learning (ML) and Discrete Choice Models (DCM) into interpretable economic frameworks. Specifically, it is concerned with the discrete representation of unobserved heterogeneity in behavioral models and how to benefit from the strengths of both DCM and ML to improve such representation without loss of economic and behavioral interpretability.

This chapter presents the motivation behind this dissertation and provides the research objectives, contributions, and outline.

### **1.1. Motivation**

Modeling and understanding human decision-making are crucial for estimating the impact of new policies or services, especially within the transportation field. All around the globe, there are concerns regarding the consequences of high levels of traffic congestion, parking demand, vehicular and greenhouse emissions, etc. Moreover, the digital revolution is reshaping every aspect of our life including the way we travel. New modes of transport, from car- and bike-sharing to Mobility on Demand (MOD) and Demand-Responsive Transit (DRT) services, are emerging as alternatives or feeders to classic public transportation systems with fixed routes and timetables. In addition, given the rapid growth rate at which the motor industry and its relevant technologies are evolving, autonomous and connected vehicles are expected to become commercially available in the near future. Predicting the impacts of such new modes on travel demand

and mobility patterns is of utmost concern to researchers, transportation planners, policymakers, and operators alike. Furthermore, changing people's travel behavior towards more sustainable transportation modes is crucial to mitigate the negative impacts of the transportation system. In behavioral science, there is a distinction between theories of change and models of behavior. While theories of change show how behaviors can change over time, models of behavior help in identifying and understanding the underlying factors that affect the prediction and explanation of a specific behavior (Darnton, 2008). Both approaches have different yet complementary objectives. Indeed, changing any kind of behavior requires a thorough investigation of all the underlying factors that lead to the behavior or decision under investigation. As for the transportation sector and along the same lines of behavioral science, modeling behavioral patterns of commuters and their decision-making process is crucial to develop sustainable and effective transport policies, predict and forecast the travel mode choices of a certain population with respect to changes in some attributes or components of the transportation system (Bhat & Lawton, 2000), and determine the different sources of taste and preference heterogeneity (El Zarwi, 2017b).

### ***1.1.1. Discrete Choice Models***

Modeling and forecasting the demand for goods or services (e.g., travel modes) from a finite set of discrete alternatives are usually conducted through Discrete Choice Models (DCM), such as the multinomial logit model (MNL) (McFadden, 1974) and its variants, which are rooted in the traditional microeconomic theory of consumer behavior and random utility maximization (Bierlaire & Lurkin, 2017). These models assume that each decision-maker associates a utility to each available alternative and

then selects the one that maximizes his/her utility. The utility of an alternative is usually specified as a linear-in-the-parameters function of, but not limited to, the alternative attributes and socio-economic/demographic characteristics of the decision-maker, in addition to a random term that represents the effect of unobserved variables. Such models are known as “explanatory” models that target behavioral and economic interpretability. However, such explanatory models encounter some limitations and might not guarantee high prediction accuracy (Sifringer et al., 2020).

Over the years, several advanced discrete choice models have been developed to overcome different problems such as the limitations of MNL, representation of unobserved taste/preference heterogeneity, and endogeneity, to name a few. The question of how best to model unobserved heterogeneity remains one of the most active research areas within demand modeling (Vij & Krueger, 2017). The mixed logit family, where choice probabilities are weighted average of standard logit probabilities over some mixing distribution (Train, 2009), is by far the most popular approach for capturing random heterogeneity. The literature is rich with studies and information on different types of mixing distributions (Yuan et al., 2015) with the main two categories being continuous and discrete. The former category assumes continuous distribution(s) with predefined forms (e.g., normal or lognormal) for the random parameters and can approximate any choice situation to a high degree of accuracy (McFadden & Train, 2000). However, such models are constrained by the predefined forms and the choice of a proper distribution which can be a complicated and computationally expensive task (Train, 2016; Vij & Krueger, 2017). Instead, discrete nonparametric representation of unobserved random heterogeneity offers an alternative perspective by making fewer statistical assumptions concerning the distributions’ forms and eliminating the

problematic and time-consuming task of choosing the right parameters' distributions. The Latent Class Choice Model (LCCM) remains the most famous and well-established example of discrete nonparametric mixing distribution and can be described as a mixed logit model with a finite mixing distribution (Train, 2008; Yuan et al., 2015). The LCCM is a random-utility model that is used whenever the modeler hypothesizes that the unobserved heterogeneity can be identified through discrete segments (or latent classes) of people that differ behaviorally from each other due to varying tastes, different decision protocols adopted by individuals and/or different choice sets considered by each individual (Gopinath, 1995). The LCCM consists of two sub-models, a class membership model that formulates the probability of an individual belonging to a specific segment/class and a class-specific choice model that estimates the choice probabilities. However, the linear-in-parameters utility specification of the latent classes may oversimplify and underestimate the extent of behavioral heterogeneity within a population (Vij & Krueger, 2017).

### ***1.1.2. Machine Learning in Choice Modeling***

Recently, due to the availability of advanced computer hardware and big data from mobile phones, social networks, and Internet-of-things, several studies have tried to apply machine learning (ML) algorithms to different transportation research areas such as traffic control, incident detection, traffic forecasting and mode choice modeling (Andrade et al., 2006; Hillel et al., 2020; Lee et al., 2018; Liang et al., 2018; Xie et al., 2003;, to name a few). Machine learning algorithms, known as “predictive” models, are non/parametric approaches that try to learn from the data without imposing strict statistical assumptions and can be used to capture complex and unobserved patterns

such as taste and preference heterogeneity. Their main target is achieving high classification and prediction accuracy (e.g., prediction of transportation modes used) rather than behavioral interpretability. Such methods, unlike discrete choice models, cannot be used to directly infer marginal effects and economic indicators such as elasticities, willingness to pay, and consumer welfare measures, which are important measures used in transportation policy and project evaluation. This lack of straightforward interpretability and the missing link with economic theories (Brathwaite et al., 2017) are believed to be the main reasons that kept the choice modeling community from trusting machine learning.

Given the different nature and purpose of DCM and ML, could we combine the strengths of both fields into hybrid frameworks in order to improve model prediction capabilities and model flexibility in representing unobserved heterogeneity without lessening the behavioral and economic interpretability? Could such hybrid frameworks be used by policymakers and transportation planners/operators for planning and policy analysis?

## **1.2. Research Objectives**

The objective of this dissertation is to integrate different machine learning algorithms within discrete choice models without compromising the behavioral and economic interpretability of the choice models. This would create hybrid models that benefit from the predictive and explanatory powers of ML and DCM, respectively. Specifically, this dissertation brings machine learning into the Latent Class Choice Model (LCCM) structure to allow for more complex and flexible discrete representation of heterogeneity, which in turn we hypothesize would result in improving the overall

model fit and prediction power. The developed models would be consistent with McFadden's four step principles of an appropriate econometric model (Manski, 2001; McFadden, 1974):

- i. The models should be consistent with random utility theory to ensure behavioral interpretability
- ii. The models must be able to forecast decision-makers' choices under different/new conditions and/or in different populations
- iii. The models should account for the fact that some attributes of the alternatives and characteristics of the decision-makers may be missing from the data in-hand
- iv. The models should be computationally practical

This dissertation focuses on the

This dissertation has three specific objectives:

- Embed Gaussian-Bernoulli mixture models, a model-based parametric clustering technique, in the traditional LCCM and investigate the impact of such practice on goodness-of-fit measures and out-of-sample prediction accuracy of the choice models.
- Embed Gaussian Processes, a nonparametric clustering technique, in the traditional LCCM and compare the resulting model to both the LCCM with mixture models and traditional LCCM in terms of interpretability, goodness-of-fit and out-of-sample prediction performance.
- Explore the forecasting performance of the proposed models and whether they lead to reasonable and/or different policy implications as compared to the LCCM.

### **1.3. Research Contributions**

Integration of machine learning and discrete choice models is still relatively a recent area of research. Apart from the contrast studies that have compared the two fields, mostly in terms of prediction performance, there are few studies that have tried to combine the two fields simultaneously into econometric frameworks. This dissertation contributes to the existing body of literature on choice modeling by presenting two new models that enhance the traditional LCCM with two different machine learning techniques. The first model makes use of a model-based parametric clustering algorithm while the second model relies on Gaussian Processes (GPs), a popular nonparametric class of probabilistic machine learning (Rasmussen & Williams, 2006).

#### ***1.3.1. Gaussian-Bernoulli Mixture – Latent Class Choice Model (GBM-LCCM)***

The first model formulates the class membership component of LCCMs as a mixture model, a method commonly used as a parametric probabilistic clustering technique in the machine learning community, to allow for more complex and flexible representation of unobserved heterogeneity. We hypothesize that this added flexibility may improve the goodness-of-fit and out-of-sample generalization (e.g., prediction accuracy) of the choice models. Specifically, the probability of a decision-maker belonging to a specific latent class/cluster is formulated as a mixture model with Gaussian and Bernoulli distributions instead of a random utility formulation.

#### ***1.3.2. Gaussian Process – Latent Class Choice Model (GP-LCCM)***

Compared to other machine learning methods, Gaussian Processes are considered more attractive due to their flexible nonparametric nature and their

formulation in a full Bayesian framework, which guarantees probabilistic interpretation of the model outputs (Mackay, 1997). Moreover, GPs are kernel-based algorithms that assume priors over latent functions rather than priors directly over parameters, which makes GPs very powerful in addressing difficult nonlinear regression and classifications problems (Rasmussen & Williams, 2006; Seeger, 2004).

Given the aforementioned advantages, the second model makes use of Gaussian Processes to replace the class membership component of the traditional LCCM. The proposed model would rely on GPs as a nonparametric probabilistic segmentation component to probabilistically divide the population into behaviorally homogenous classes while simultaneously relying on random utility models to develop class-specific choice models. To the author's knowledge, this dissertation formulates the first Gaussian Process choice model within an LCMM framework, thereby allowing for more modeling flexibility and potentially higher prediction accuracy.

### ***1.3.3. Estimation and Application***

Moreover, this dissertation provides the formulation and implementation of two Expectation-Maximization (EM) based algorithms for the estimation of the two proposed models. The two algorithms would benefit from the EM iterative nature to jointly estimate/infer the hyper/parameters of the machine learning algorithms and the class-specific choice parameters. Using different mode choice applications, the proposed models are compared to their traditional LCCM counterpart on the basis of parameter estimate signs, value of travel time savings, statistical goodness-of-fit measures, and cross-validation tests.



## **1.4. Outline**

This dissertation is structured as follows.

Chapter 2 presents the necessary background material. It starts by reviewing discrete choice models, the concept of taste heterogeneity and the problems facing its representation in behavioral models. It then reviews studies that have used machine learning techniques in travel mode choice modeling and the ones that have tried to combine machine learning with econometric models. Throughout, it highlights the limitations of both machine learning and discrete choice models to motivate the need for hybrid frameworks that combine the two fields.

Chapter 3 provides the model structure of the proposed GBM-LCCM. It also provides the formulation and derivation of the corresponding Expectation-Maximization algorithm.

Chapter 4 presents two mode choice applications to assess the proposed GBM-LCCM and compare it with different benchmark models.

Chapter 5 provides the model structure of the second proposed model (GP-LCCM), as well as the formulation and derivation of the corresponding Expectation-Maximization algorithm.

Chapter 6 presents three mode choice applications to assess the proposed GP-LCCM and compare it with different benchmark models in addition to the proposed GBM-LCCM.

Chapter 7 comprises the development of a policy analysis using the American University of Beirut case study. It presents and compares the forecasting results of the two proposed models, GBM-LCCM and GP-LCCM, as well as the traditional LCCM.

Chapter 8 concludes the dissertation. It provides a comprehensive summary of the research objectives, research contributions, proposed models, and findings. It then discusses limitations of this dissertation and future research directions.

## CHAPTER 2

### LITERATURE REVIEW

As previously discussed in Chapter 1, this dissertation attempts to improve the discrete representation of heterogeneity in behavioral models by combining discrete choice models and machine learning techniques. Therefore, this chapter presents the corresponding background material. It starts by reviewing discrete choice models based on random utility theory including the MNL formulation and its criticisms (Section 2.1.1). It then discusses the concept of taste heterogeneity and the problems facing its representation in behavioral models (Section 2.1.2). Next, this chapter reviews the use of machine learning in choice modeling (Section 2.2). It presents the different aspects of machine learning (Section 2.2.1), studies that have used machine learning techniques in travel demand models (Section 2.2.2), and differences between machine learning and discrete choice models (Section 2.2.3). This chapter then reviews studies that have developed hybrid frameworks to combine the two fields (Section 2.3). Throughout, it discusses the advantages and disadvantages of each field in order to motivate the need for the hybrid models proposed by this dissertation (Section 2.4).

#### **2.1. Discrete Choice Model**

Econometric discrete choice models, derived from random utility maximization theory, have been widely used to model choices made by decision-makers among a finite set of discrete alternatives. These models are used in different fields such as transportation, economics, finance, marketing, medicine, psychology etc. Some studies tried to replace the theory of utility maximization from discrete choice models with

different behavioral concepts (Chorus, 2012). For instance, the random regret minimization model replaces the utility maximization theory with regret theory and as such assumes that an individual tries to minimize anticipated regret when choosing between alternatives instead of maximizing utility (Chorus et al., 2008). Although random regret minimization models allow for the possibility that choices might be driven by the wish to minimize regret or negative emotions, they violate, as opposed to random utility models, several microeconomic axioms that ensure complete economic and welfare analysis (e.g., disaggregate elasticities, willingness to pay, consumer welfare measures<sup>1</sup>) (Chorus, 2012; Dekker & Chorus, 2018).

### ***2.1.1. MNL Formulation***

Early forms of random utility maximization models were developed during the 1960s by Marschak (1960) and Cox (1966). However, it was McFadden's contribution to discrete choice analysis during the 1970s, the conditional logit model (1974), that received more attention from econometricians and researchers (Brathwaite et al., 2017; Manski, 2001). This is mainly due to the fact that he linked his MNL formulation to the classical consumer demand theory (McFadden, 2001). According to McFadden's formulation, any econometric behavioral model should fulfil four main properties (Manski, 2001). First, the model should be consistent with utility theory, meaning that a decision-maker  $n$  facing a finite set of alternatives would select the alternative that

---

<sup>1</sup> **Disaggregate elasticity** is the change in the choice probability of an individual due to a change in the level of some attribute (Ben-Akiva & Lerman, 1985). **Willingness to pay** is the maximum amount of money a consumer is willing to spend on a service or good and is calculated as the ratio between the marginal utility of an attribute and the marginal utility of cost (e.g., value of time is the marginal utility of travel time divided by the marginal utility of travel cost) (Ben-Akiva & Lerman, 1985). **Consumer welfare measures** are the benefits an individual obtains from the consumption of goods or services. A common measure in transport demand modeling is consumer surplus or logsum which is equal to the utility, in monetary terms, that an individual receives in the choice situation (Train, 2009).

maximizes his/her utility. Second, researchers must be able to forecast decision-makers' choices under different/new conditions and/or in different populations. This is achieved by defining the utility  $U_{nj}$ , that a decision-maker  $n$  might gain from choosing alternative  $j$ , as a function of some observed attributes of alternative  $j$  ( $X_{nj}$ ) and characteristics of decision-maker  $n$  ( $S_n$ ). Third, the econometric analysis should account for the fact that the researcher will not be able to observe all aspects of the utility. Typically, some attributes of the alternatives and characteristics of the decision-makers will be missing from the data in-hand. Therefore, utility  $U_{nj}$  is decomposed into two parts, a systematic utility  $V_{nj}$  and a random disturbance term  $\varepsilon_{nj}$ :

$$U_{nj} = V_{nj} + \varepsilon_{nj} , \quad (1)$$

$$V_{nj} = \beta_1 X_{nj} + \beta_2 S_n . \quad (2)$$

The systematic utility  $V_{nj}$ , also known as representative utility, relates the observed components  $X_{nj}$  and  $S_n$  to two vectors of unknown parameters,  $\beta_1$  and  $\beta_2$ , that need to be estimated statistically using the available data. The disturbance  $\varepsilon_{nj}$ , a random term with a specific density hypothesized by the modeler, accounts for the contribution of the unobserved factors. Once the probability distribution of  $\varepsilon_{nj}$  is specified, the researcher can estimate the probabilities of the decision-makers' choices. Finally, the econometric model should be computationally practical (Manski, 2001). Given the available technology during the 1970s, McFadden used a simple closed form for the conditional choice probability by assuming that the unobserved utility terms are independently, identically distributed (over individuals and alternatives) as Extreme Value Type I. This assumption allowed McFadden to develop the famous multinomial logit formulation (McFadden, 1974):

$$P_{nj} = \frac{e^{V_{nj}}}{\sum_{j'=1}^J e^{V_{nj'}}}, \quad (3)$$

where  $P_{nj}$  is the logit probability of choosing alternative  $j$  from  $J$  available alternatives.

Researchers have trusted this MNL formulation due to its connection to consumer theory, closed form choice probabilities and simple interpretability. However, logit models suffer sometimes from strict statistical assumptions, such as the independence of irrelevant alternatives (IIA) which leads to proportional substitution patterns across alternatives (Train, 2009). While the IIA assumption captures people's behavior accurately in some situations, it might generate biased demand estimates in many other applications. In addition, the logit model can only represent taste variations (differences in choice behavior among individuals) when heterogeneity in the choice process varies systematically and not randomly and can only deal with panel data (i.e., data collected from the same individuals over time) when unobserved factors are uncorrelated over time and individuals.

During the last decades, different advanced discrete choice models have been developed to relax the behavioral limitations of the MNL model while concurrently satisfying the above four properties. Generalized Extreme Value (GEV) models (e.g., nested logit) relax the IIA assumption allowing for flexible substitution patterns. Probit models overcome the three main limitations of the logit but their disturbance terms are restricted to normal distributions while mixed logit has better flexibility and can embed any distribution (normal, log-normal, triangular, etc.) for the unobserved factors and/or the coefficients of the observed attributes (Train, 2009). Although these advanced models have higher flexibility and can better predict people's choices, they still encounter difficulties when dealing with complex datasets with high degrees of nonlinear relationships between variables (Karlaftis & Vlahogianni, 2011; Lee et al.,

2018). Moreover, the field of discrete choice modeling still struggles with the question of how to better represent heterogeneity in the choice process (Vij & Krueger, 2017).

### 2.1.2. Taste Heterogeneity

#### Systematic/Random Specifications

Heterogeneity is known as taste variation across decision-makers (i.e., different people have different sensitivities to the same attribute) and is usually captured through systematic or random specifications. When tastes vary systematically with observable variables, heterogeneity in the choice process is represented through interactions between socioeconomic characteristics related to the decision-makers and attributes of the alternatives. However, systematic specifications can lead to false conclusions, unreliable parameter estimates, and incorrect forecasts in case tastes vary randomly across decision-makers or are related to unobserved variables (Gopinath, 1995; Vij et al., 2013). Random taste heterogeneity is typically captured through mixed logit models which can approximate any random utility model (McFadden & Train, 2000). Mixed logit probabilities are defined as a weighted average of standard logit probabilities evaluated over a mixing distribution (density) of parameters. The probability of individual  $n$  choosing alternative  $j$  can then be expressed as follows:

$$P_{nj} = \int \frac{e^{V_{nj}}}{\sum_{j'=1}^J e^{V_{nj'}}} f(\beta) d\beta, \quad (4)$$

$$V_{nj} = \beta_{n1}X_{nj} + \beta_{n2}S_n, \quad (5)$$

where  $\frac{e^{V_{nj}}}{\sum_{j'=1}^J e^{V_{nj'}}$  is the standard logit choice probability and  $\beta_{1n}$  and  $\beta_{2n}$  are two vectors of unknown coefficients that vary over decision makers with a predefined density  $f(\beta)$ .

This specification allows for different tastes/coefficients within the population. Most mixing distributions fall typically under two categories: parametric (also known as continuous mixed logit) and nonparametric distributions. Parametric distributions have predefined forms (e.g., normal, lognormal, etc.) with fixed parameters and usually provide great fit to the data. However, the choice of a proper distribution can be complicated and computationally expensive. Researchers have to make a prior assumption about the proper distribution or estimate different models with different distributions and then choose the best model based on statistical goodness-of-fit measures and behavioral interpretation of the parameter estimates (Vij & Krueger, 2017). Moreover, parametric distributions have limited flexibility due to their predefined shapes that can be classified as bounded or unbounded. Bounded distributions can be adopted when the analyst has an a priori belief of obtaining an explicit sign for a specific coefficient. The most known bounded distribution is the lognormal, which has been adopted in some studies with great success (Bhat, 1998, 2000; Train & Sonnier, 2004). However, the lognormal distribution can increase the estimation time significantly and overestimate the mean and standard deviation due to its long tail property (Hess and Polak, 2004; Hess et al., 2005). The triangular distribution can avoid the issues of long tails of the lognormal distribution and the symmetrical shape of the normal distribution. However, it is rarely used due to its simple linear shape. Unbounded distributions (e.g., normal) can also be used since restricting some coefficients to be strictly positive or negative by using distributions with fixed bound at zero can limit the ability of the model to reveal some counter-intuitive information contained in the dataset and lead to poorer model fit (Hess et al., 2005). Several other parametric forms have been used in the literature (e.g., truncated



normal, Johnson's  $S_B$  etc.) but the choice of a proper distribution and the shapes that these distributions can fit are still considered as two major drawbacks in discrete choice modeling. Furthermore, most of the parametric models estimated in the literature are limited to univariate distributions although some studies have tried to use mixture of continuous distributions as a random taste parameter distribution (e.g., Fosgerau and Hess, 2009; Keane and Wasi, 2013).

### Nonparametric Distributions

To overcome these constraints, researchers have relied on nonparametric distributions which do not have predefined shapes and do not require the researcher to make certain assumptions regarding the distributions of parameters across decision-makers, meaning more flexibility can be guaranteed (Yuan et al., 2015). The LCCM remains the most known and used nonparametric distribution. It is a random utility model that extends the multinomial logit model by using the concept of latent class formulation and allows capturing heterogeneity in the choice process by allocating people probabilistically to a set of  $K$  homogeneous classes that differ behaviorally from each other. It is typically used when the modeler postulates that the unobserved heterogeneity can be represented by discrete latent classes such as segments of the population with varying tastes, different decision protocols adopted by individuals, and choice sets considered which may vary from one individual to another (Gopinath, 1995).

Such model consists of two sub-components, a class membership model and a class-specific choice model (Figure 1). The class membership model formulates the probability of a decision-maker belonging to a specific class, typically as a function of his/her characteristics. Conditioned on the class membership of the decision-maker, the

class-specific choice model estimates the probability of choosing a specific alternative as a function of the observed exogenous attributes of the alternatives and characteristics of the decision-maker. This framework divides or segments decision-makers into homogeneous groups through a probabilistic model that uses observed exogenous variables as input.

It is to be noted that any model that combines discrete choice models with continuous and/or discrete latent variable models, such as the LCCM, is considered as part of the Hybrid Choice Model (HCM) family (Abou-Zeid & Ben-Akiva, 2014; Ben-Akiva, McFadden, et al., 2002).

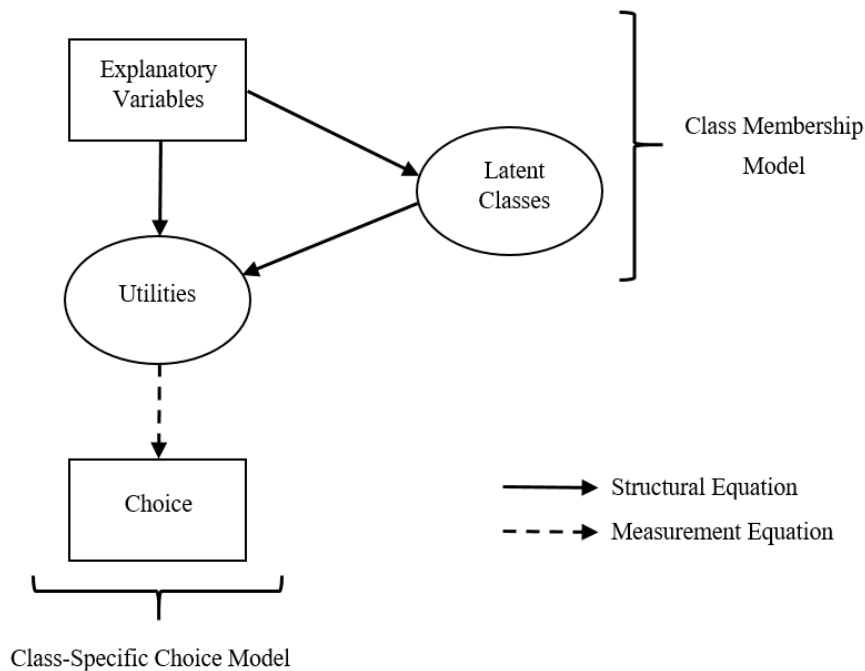


Figure 1: Latent Class Choice Model Framework (adapted from Walker and Li (2007))

### LCCM vs Continuous Mixed Logit

Several studies have tried to compare both continuous mixed logit and LCCM from theoretical and empirical perspectives (Andrews et al., 2002; Greene and Hensher, 2003; Han, 2019; Hess et al., 2009, to name a few). To sum up, LCCM has some

advantages over the continuous mixed logit. First, LCCM makes fewer statistical assumptions regarding the parameters' distribution form. Second, unobserved heterogeneity in continuous mixed logit models suffers from a lack of interpretability since it is not usually explained by explanatory variables, although it is possible (Greene et al., 2006), while discrete latent classes are easily explained and interpreted since the class membership model of LCCM is usually a function of socio-economic characteristics of decision-makers. Third, correlation between taste parameters and elasticities are two major differences between the two approaches. In continuous mixed logit models, correlation can be accounted for by specifying a joint distribution for taste parameters; however, most applications rely on independently distributed random taste parameters. As for LCCM, correlation between taste parameters is implicit in the model and it is a function of the class membership probabilities, which are a function of the socio-economic variables, and the class-specific taste parameters. The same rationale applies to the relationships between the elasticities and the socio-economic variables which are not easily determined in continuous mixed logit models but are directly inferred from the class membership probabilities of LCCM (Hess et al., 2009). One major shortcoming of LCCM is that the discrete latent representation may oversimplify the unobserved heterogeneity, especially when a small number of classes is estimated, since latent classes are defined as a linear-in-the-parameters function of the socio-economic characteristics of the decision-makers. In general, the flexibility of LCCM increases with the number of classes. However, the computational complexity grows rapidly since the number of parameters increases with the number of classes as well (Yuan et al., 2015). Consequently, the computational burden that precludes the estimation of LCCMs with a high number of classes in addition to the linear-in-

parameters specification of the latent classes may generate in practice simpler (less flexible) models than the LCCM framework can offer (Vij & Krueger, 2017).

In order to loosen some of the restrictions of continuous mixed logit and LCCM, several studies have relied on mixture of distributions approaches. For example, Bujosa et al. (2010) developed a Latent Class-Random Parameter Logit model (LC-RPL) to combine the concepts of latent classes and random taste parameters. The model outperformed the traditional LCCM and continuous mixed logit models in terms of goodness-of-fit and in-sample predictions. However, the application was limited to two latent classes and a univariate normal distribution for one taste parameter. A similar approach was implemented by Greene and Hensher (2013). The proposed model had better goodness-of-fit measures than the traditional LCCM and continuous mixed logit model but was also limited to two latent classes and one univariate triangular distribution for a taste parameter. Fosgerau and Hess (2009) compared two mixture approaches against four continuous mixed logit models with different continuous distribution functions (normal, lognormal, triangular, and SB). The first approach uses a Mixture of Distributions (MOD) to define the distributions of random taste parameters while the second one uses the Normal distribution as a base for the random parameters and extends it by adding a series approximation of Legendre polynomials. The MOD approach had a slight advantage over the second approach and the traditional mixed logit models. However, it had computational problems and it was not possible to estimate more than a mixture of two normal distributions. Krueger et al. (2018) presented a Dirichlet process mixture multinomial logit (DPM-MNL) model where Dirichlet process is used as a flexible mixing distribution for the parameters. Such approach does not require the analyst to specify the number of mixtures a priori.

However, it generates unstructured representations of heterogeneity which affects the interpretability of the model. Train (2008) developed an Expectation-Maximization (EM) algorithm for the estimation of mixture of distributions in mixed logit models. However, the application was also limited to mixture of two independent distributions for each randomly distributed parameter. Moreover, Train (2016) introduced a new logit-mixed logit model where he relied on logit specifications to define the mixing distribution of random parameters. The framework proved its capability to approximate the shape of any mixing distribution but placed additional burden on the analyst to specify the utility of the random parameters and the variables that represent the shape of their distributions.

## **2.2. Machine Learning in Choice Modeling**

In recent years, the use of machine learning techniques has witnessed a major growth due to the exponentiation in the amount of data available (e.g., from mobile phones, social networks, internet-of-things, etc.), as well as incredible computing resources advances. Such methods are being applied to problems from different fields (speech processing, computational biology, finance, robotics, computer vision, natural language processing, etc.). As for transportation, researchers have been exploring the feasibility of applying machine learning techniques to different transportation research areas such as traffic control (Abdulhai et al., 2003; Bingham, 2001; Srinivasan et al., 2006), incident detection (Jin et al., 2002; Srinivasan et al., 2004; Wang et al., 2008), traffic forecasting (Deshpande & Bajaj, 2017; Hong et al., 2011; Ma et al., 2018; Stathopoulos et al., 2008; Vlahogianni et al., 2008), prediction of transportation modes from raw GPS data and/or mobile phone sensors such as accelerometers and gyroscopes

(Dabiri & Heaslip, 2018; Gonzalez et al., 2010; Jahangiri & Rakha, 2015, 2014; X. Zhu et al., 2017) etc. Before discussing machine learning in mode choice modeling, we start by presenting an overview of machine learning and its main categories (Section 2.2.1). Next, we review comparative studies between machine learning and discrete choice models (Section 2.2.2) and discuss the differences between the two fields (Section 2.2.3).

### ***2.2.1. General Description***

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that is concerned with developing algorithms capable of finding or “learning” patterns in empirical data (Wittek, 2014) and generalizing well (maximizing their predictive accuracy) on unseen data that was not used for training/estimation. Machine Learning algorithms can be divided, according to the “learning” type, to three major categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning requires labeled data to guide the learning algorithm. Labeled data means input data that has both independent (explanatory) variables/features and dependent variables (classes, targets, outputs, or labels) whose values need to be estimated and predicted. A supervised algorithm makes use of a labeled sample to build a classifier that would assign labels to both training and test samples. Training sample or dataset is the data used for training/estimating the model while test sample or dataset is a data that was not used during the training process and whose labels have to be predicted by the model.

The supervised learning subcategory of ML includes several algorithms such as Decision Trees (DTs), Support Vector Machines (SVMs), Artificial Neural Networks (ANNs) etc. According to the type of labels, supervised problems can be divided into

two main sub-categories, classification and regression. In classification problems, the goal is to predict discrete labels (e.g., true or false, spam or not spam, travel modes) while the goal of regression problems is to predict continuous variables (e.g., home prices, vehicle miles traveled). In unsupervised settings, the labels or dependent variables are usually missing and the ML algorithms make use of only independent variables/features to identify underlying patterns or structures from an unlabeled dataset. The three main sub-categories of unsupervised machine learning are clustering, density estimation, and dimensionality reduction. Clustering methods are used to discover clusters (groups or classes) of similar characteristics within the data, density estimation methods focus on determining the distribution of data, and dimensionality reduction techniques project the data from a high dimensional space to lower-dimensional representations of data. Finally, the third main category of machine learning, reinforcement learning, relies on rewards to discover the best decision or behavior an algorithm should perform.

In ML terminology, discrete choice models for mode choice modeling can be considered as a supervised classification technique (Hillel et al., 2020) since the goal is to model decision-makers' choices among a finite set of discrete alternatives (i.e., classes), given a set of independent variables (i.e., features). However, in this dissertation, we focus on unsupervised learning, mainly clustering, for reasons we discuss later.

### ***2.2.2. Comparative Studies***

Machine learning techniques, mainly supervised classification algorithms, are increasingly being used in mode choice modeling as alternative methods to traditional

econometric models. Nijkamp et al. (1996) explored the modal split between rail and road transport modes using Artificial Neural Networks and MNL models. Results showed that ANN has slightly better generalization performance (e.g., prediction accuracy) than the logit model. Xie et al. (2003) compared Decision Trees and Artificial Neural Networks with MNL models in the context of commuter mode choice. They showed that both DT and ANN provide better prediction accuracy. Vythoulkas and Koutsopoulos (2003) used fuzzy sets and concluded that MNL has a slightly worse rate of accurate predictions. Artificial Neural Networks have been also applied by Cantarella and de Luca (2005) to two case studies with different trip purposes (work vs. educational trips). Results showed that ANN has a significant edge over several econometric models (MNL, Nested Logit, and Cross-Nested Logit) especially when the mode shares are similar. Neural networks with neuro-fuzzy inference systems have been also used by Andrade et al. (2006) for shopping mode choice modeling and results demonstrated better predictive performance as compared to an MNL model. Several other studies have also used neural networks with different architectures in the context of travel mode choice analysis and showed that better generalization performance results are achieved compared to MNL and/or nested logit models (e.g., Hagenauer and Helbich, 2017; Lee et al., 2018; Omrani, 2015; Omrani et al., 2013; Xian-Yu, 2011; Zhang and Xie, 2008).

Support vector machines have been also applied for travel mode choice applications. For instance, Zhang and Xie (2008) compared SVM, ANN, and MNL while Xian-Yu (2011) compared SVM, ANN and nested logit. Both studies found that SVM outperforms ANN and logit models in terms of prediction accuracy. On the other



hand, other studies (Omrani, 2015; Omrani et al., 2013) found that ANNs have higher accuracy than SVM and MNL models.

Decision Trees and its variants have been also adopted by several studies. Xie et al. (2003) showed that DTs are more efficient than ANNs and guarantee some level of interpretability due to the if-then rules used while constructing the trees. People's mode switching behavior when the choice is restricted between two modes only was investigated by Tang et al. (2015) using DT and MNL models. It was found that DT models achieve higher prediction accuracy especially in detecting switching to minority modes. Liang et al. (2018) chose Random Forest (RF), an ensemble of randomly constructed decision trees, to estimate households' travel mode choices and compared the results with an MNL model using different sample sizes. First, the results showed that the prediction accuracy of the RF model was slightly higher. Second, both models achieved the highest accuracy with a sample size between 2,000 and 6,000 observations. Finally, the accuracies of MNL and RF models fluctuated unsteadily with small samples and decreased as the sample size increased. Sekhar et al. (2016) showed the superior predictive capability of an RF (98.96%) to an MNL model (77.31%). Moreover, Hagenauer and Helbich (2017) compared the predictive performance of an MNL model with six machine learning techniques, Naïve Bayes (NB), ANN, SVM, and tree-based ensemble methods including Boosting, Bagging, and Random Forest. The study showed that MNL had the lowest prediction accuracy while tree-based ensemble techniques provided the highest accuracies with RF performing significantly better than Bagging and Boosting classifiers. Wang and Ross (2018) also used an ensemble tree-based classifier called Extreme Gradient Boost (XGB). They showed that both XGB and MNL models performed poorly when predicting the choice of modes with small

shares. However, the XGB model significantly outperformed the MNL model in terms of prediction accuracy when the dataset is balanced.

The literature is rich with many other studies that have compared discrete choice models and supervised machine learning algorithms, particularly classification techniques, in the context of mode choice modeling. For a more comprehensive and exhaustive review of such comparative studies, readers may refer to the following review papers (Hillel et al., 2020; Minal & Sekhar, 2014; Ratrout et al., 2014). The first two papers (Minal & Sekhar, 2014; Ratrout et al., 2014) reviewed studies that have used ANNs for mode choice modeling problems and concluded that ANNs are successful in such applications due to their flexibility and ability to handle large and nonlinear datasets. The third and most recent review paper (Hillel et al., 2020) reviewed 73 studies that have used supervised machine learning algorithms to investigate passenger mode choice, including: Logistic Regression (LR), Artificial Neural Networks, Support Vector Machines, Decision Trees, and Ensemble Learning (EL). The study shows that half of the reviewed articles were published after 2014 while only 14% of the papers were published prior to 2007. This highlights the growing trend of using supervised machine learning techniques for mode choice modeling in the past few years and indicates that such trends are expected to continue in the future. Hillel et al. (2020) also argues that machine learning techniques do not provide straightforward behavioral interpretability and economic indicators to inform policy making decisions. Out of the 73 reviewed studies, only four (Andrade et al., 2006; Ding et al., 2018; Subba Rao et al., 1998; S. Wang & Zhao, 2019) tried to extract behavioral indicators (e.g., aggregate point elasticities, VOT) by performing sensitivity or elasticity analysis while only five other studies (Errampalli et al., 2007; Kedia et al., 2015; Kumar et al., 2013; Lee et al.,

2018; Pulugurta et al., 2013) estimated aggregate mode shares for different policy scenarios by changing the variables of interest and re-estimating the models.

A few studies in choice modeling considered applications other than travel mode choice. For instance, Mohammadian and Miller (2002) showed that ANN outperforms nested logit in terms of prediction of household automobile choices. Biagioni et al. (2009) developed an ensemble of conditional and unconditional classifiers to explore the mode choice at the tour level. First, the unconditional model predicts the mode of the first trip in a tour, then the conditional model makes predictions for the remaining trips. The best predictive performance was achieved by using Naïve Bayes and Decision Trees for the unconditional and conditional classifiers, respectively. Results showed that this two-step framework outperforms the traditional MNL model in terms of prediction accuracy. Golshani et al. (2018) compared the performance of ANNs with discrete, continuous, and discrete-continuous statistical models by modeling travel mode and departure time choices. Accelerated hazard model was used to model trip departure time as a continuous variable while MNL was used for travel mode choice modeling. Moreover, the copula-based approach (Bhat & Eluru, 2009) was employed to jointly model the discrete (travel mode) and continuous decisions (departure time). It was found that the ANN models have a significant edge over the three statistical models in terms of prediction accuracy and implementation.

Contrary to the above-mentioned literature, few contrast studies have shown no clear advantage of machine learning over econometric models in terms of prediction power. For instance, Sayed and Razavi (2000) compared ANNs with MNL and MNP in the context of freight transport mode choice and reported similar accuracy for the three

approaches. Similarly, Hensher and Ton (2000) found no clear superior performance for ANN over nested logit.

Beyond the comparison between the two fields and the use of supervised machine learning techniques, some recent studies have relied on generative (unsupervised) methods. For instance, Van Cranenburgh and Alwosheel (2019) investigated decision rule heterogeneity among decision-makers using deep learning networks while Wong et al. (2018) used a restricted Boltzmann machine to estimate latent (or unobserved) variables without relying on measurement indicators of the latent variables and attitudinal questions which are typically used in behavioral studies and hybrid discrete choice models (or Integrated Choice and Latent Variable (ICLV) models).

### ***2.2.3. Differences***

Although most of the aforementioned machine learning techniques have shown a superior prediction accuracy compared to econometric discrete choice models, econometricians and transportation researchers are still relying on traditional econometric models instead of machine learning algorithms. Moreover, researchers from both disciplines usually fail to communicate (Karlaftis & Vlahogianni, 2011) and make the most of both approaches. This may be due to the differences in the terminology, philosophy and goals, model evaluation, and assumptions/limitations of the two approaches as further explained below (Karlaftis & Vlahogianni, 2011).

#### **2.2.3.1. Terminology**

First, although some studies have made an effort to bridge the gap between the different concepts and terms used in machine learning and econometric models (Sarle,

1994), terminology remains a main source of confusion for researchers when trying to compare or relate the two approaches (Table 1).

Table 1: Equivalent terminology between Machine Learning and Econometric Models (adapted from Sarle (1994); Karlaftis and Vlahogianni (2011); Hillel et al., (2020))

<b>Machine Learning</b>	<b>Econometric Models</b>
Targets, outputs, labels	Dependent variables
Clusters	Classes
Classes	Alternatives
Input, features	Independent variables, attributes
Errors	Residuals
Training	Estimation
Error or cost function	Estimation criterion
Weights	Parameter estimates or coefficients
Functional links	Transformations
Intercept/bias	Alternative Specific Constant (ASC)

### 2.2.3.2. Philosophy and Goals

Second, the main difference lies in the underlying philosophy and goals of the two approaches. Many machine learning algorithms are usually known as “black boxes”, in which highly complex non/parametric functions are used in order to improve the model prediction accuracy. The term “black box” is used since underlying causal relationships and inference from the explanatory variables are de-emphasized. Therefore, machine learning models can be referred to as “predictive” models that target high classification and prediction accuracy at the expense of interpretability, although few studies have shown that different techniques can be applied to extract some behavioral indicators (Andrade et al., 2006; Ding et al., 2018; Wang et al., 2020; Wang & Zhao, 2019). On the other hand, traditional discrete choice models are known as “explanatory” models that assume parametric relationships between the utility (or

desirability) of each alternative and its potential attributes. They can be used to directly infer marginal effects and economic indicators such as elasticities, willingness to pay, and consumer welfare measures. However, these explanatory models might not guarantee high prediction accuracy. Moreover, traditional DCMs are rooted in microeconomic theories of human decision-making behavior (Bierlaire & Lurkin, 2017). It is believed that this connection is the main reason econometricians and transportation planners have heavily relied on discrete choice models and specifically the MNL formulation of McFadden (Brathwaite et al., 2017; McFadden, 2001). It is also believed that the main reason that kept econometricians from trusting machine learning is the missing link with economic theories (Brathwaite et al., 2017).

#### 2.2.3.3. Assumptions and Limitations

Third, another difference is in the assumptions/limitations of the two approaches. Econometric models impose a priori and often strict statistical assumptions regarding the error term while machine learning techniques are more flexible since few assumptions to none are made. Furthermore, econometric models encounter difficulties when dealing with complex datasets with high degrees of nonlinearity (Karlaftis & Vlahogianni, 2011; Lee et al., 2018), outliers or noisy data, and correlated explanatory variables.

#### 2.2.3.4. Model Evaluation

Finally, another main difference lies in the model evaluation and selection process of the best model. Machine learning models are mainly evaluated using performance (prediction) measures since the main goal is building a model with high prediction accuracy. Performance measures can be categorized as discrete and probabilistic metrics. The main discrete performance is “Confusion matrix”, a matrix

with four different combinations of predicted and actual values: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It is extensively used in computing the following discrete performance metrics: accuracy (proportion of total number of predictions that were correct), precision (proportion of predicted positive cases that were correct,  $TP/(TP + FP)$ ), recall (proportion of positive cases that were correctly identified,  $TP/(TP + FN)$ ), and F1-score (a harmonic mean of precision and recall). The main probabilistic metrics are: i) Receiver Operating Characteristic (ROC) curve, a plot of the True Positive rate against the False Positive rate at all classification thresholds; ii) Mean Absolute Error (MAE) or Mean Squared Error (MSE), the average difference between the actual observations and the predicted observations; iii) Logarithmic Loss defined as follows:  $-\frac{1}{N} \sum_{n=1}^N \sum_{j=1}^M y_{nj} \log(p_{nj})$  where  $N$  is the sample size,  $M$  is the number of classes,  $y_{nj}$  indicates whether or not observation  $n$  belongs to class  $j$ , and  $p_{nj}$  is the probability of observation  $n$  belonging to class  $j$ ; iv) Bayesian Information Criterion (BIC). Several other performance measures exist but the choice of the right one depends on the type of the application.

Further, overfitting is a main concern in machine learning. A model might perfectly fit the training data but fail to make good predictions on future unseen datasets. Therefore, what matters in model evaluation is the generalization performance, i.e. the predictive performance of the developed model on unseen (or test) data. Several techniques can be used when evaluating the generalization performance. The simplest method is the holdout technique. First, the data is randomly divided into two sets, a training set and testing set. Second, the model is fitted using the training set only. Finally, the fitted model is applied to the testing set and performance measures are estimated. In practice, the most famous and used technique is the  $k$ -fold cross-

validation. This technique requires dividing the data into  $k$  equally sized sets (or folds). The model is then trained on  $k - 1$  folds and tested on the remaining one. This process is repeated  $k$  times in order to test the generalization performance of the model on the entire data. Another technique is called leave-out method, a special case of  $k$ -fold cross validation technique where  $k$  is equal to  $N$  (number of observations).

As for econometric discrete choice models, models are usually evaluated on the basis of parameter estimate signs and magnitudes, reasonableness of estimated economic measures (e.g., willingness to pay, elasticities), statistical goodness-of-fit measures such as the likelihood ratio test, the robust t-test, and the adjusted rho-square, and through comparison of information criteria such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) etc. Econometric models can also be evaluated on their prediction capability although it is not as commonly practiced as in machine learning.

To sum up, machine learning models are usually evaluated on the basis of their predictive power while the evaluation techniques in econometric analysis have traditionally mainly focused on the explanatory power of the models. Recently, some studies have used the concept of cross validation technique to evaluate and compare the predictive power of different discrete choice models (e.g., Robin et al., 2009; Robin and Bierlaire, 2012; Sfeir et al., 2020). They randomly divided the data into a train and test set. Then, they estimated the parameters of the model with the train set and calculated a measure of fit of the estimated model, typically the log-likelihood, on the test set. Other studies from the machine learning community have used different sensitivity analysis techniques to explore the effect of explanatory variables on dependent variables (outputs) (Golshani et al., 2018; Hagenauer and Helbich, 2017; Lee et al., 2018, to name



a few). However, most of the abovementioned comparative studies (section 2.2.2) have only compared the predictive power of the two approaches. Moreover, Hillel et al. (2020) found in their review paper that 60 out of the 73 reviewed studies used only discrete metrics to evaluate model performance. Relying on discrete metrics only and neglecting probabilistic metrics during the evaluation process is risky and will most likely lead to non-representative mode shares especially in the case of imbalanced datasets since discrete metrics will always assign observations to the classes (alternatives) with the highest probabilities (Hillel et al., 2020). Furthermore, any comparative study or hybrid framework which integrates both approaches has to investigate the explanatory power as well as the predictive power (mainly probabilistic metrics) of the developed models before rushing to conclusions as to which approach is better.

### **2.3. Combining the Two Fields**

As previously mentioned, both approaches have their advantages and disadvantages. Recently, some studies have tried to connect the two fields through hybrid frameworks. For instance, a two-stage sequential logit-Artificial Neural Networks framework for choice modeling has been proposed by Gazder and Ratrouf (2016). First, they developed several logit models for different existing and hypothetical mode choice situations. Second, they trained different ANN models to predict the mode choice using the logit probabilities as input. Finally, they compared the accuracy of the integrated approach with separate logit and ANN models. Results showed that the proposed model improves the generalization performance (prediction accuracy) in multinomial choice situations while logit models have marginally higher prediction

accuracies in binary choice situations. Sifringer et al. (2018) also developed a two-stage sequential model by adding an extra term, estimated through a Deep Neural Network (DNN), to the utilities of the logit model. The extra term was estimated separately by using all disregarded variables in the logit model as input to the DNN model. This approach improved the log-likelihood of the simple logit model by around 15% without weakening the statistical significance of the logit parameters. Along the same lines, Sifringer et al. (2020) have extended the MNL and nested logit by integrating a non-linear representation arising from an ANN into the utilities to improve the specification of both logit models. Han et al. (2020) developed a TasteNet-MNL model by embedding a neural network into the utilities of a logit model to improve the systematic representation of heterogeneity. Both sub-components of the TasteNet-MNL, the neural network and the logit model, are estimated simultaneously and the overall model can be considered as an extension of the sequential model of Sifringer et al. (2018). Another choice model that combines neural networks and random utility models (Wong & Farooq, 2019) has been developed by using the concept of residual learning within a neural network architecture to allow for training of deep neural networks and as such identifying complex sources of unobserved heterogeneity.

#### **2.4. This Dissertation in Context**

Most of the previous studies have focused on applying supervised machine learning to classifications tasks such as mode choice modeling or on combining machine learning techniques (mostly Neural Networks) and discrete choice models in sequential or simultaneous frameworks. Furthermore, the majority of studies that have used machine learning methods as alternatives to discrete choice models for mode

choice modeling have mostly focused on supervised applications and prediction accuracy (Karlaftis and Vlahogianni, 2011; Wong and Farooq, 2019; Hillel et al., 2020) often at the expense of economic interpretability due to the disconnection of such techniques from economic principles and theories (Brathwaite et al., 2017). Some recent studies have however shown that machine learning techniques can provide practical economic information (Brathwaite et al., 2017; Wang et al., 2020).

Most of the studies on machine learning for mode choice modeling have been related to classification tasks while clustering techniques have not yet been rigorously addressed in the literature, although few studies have addressed this clustering aspect of machine learning. For instance, Han (2019) developed a nonlinear-LCCM by using neural networks to specify the class membership model. The proposed model with 8 latent classes outperformed the best LCCM with 6 latent classes in terms of prediction accuracy. However, the nonlinear-LCCM is less transparent and loses interpretability at the latent class level due to the “black-box” nature of neural networks.

This dissertation aims at embedding unsupervised machine learning (clustering) in an econometric framework that satisfies McFadden’s vision of a proper choice model. Clustering methods are used to discover heterogeneous subgroups or latent classes within a population by allocating similar observations (e.g., individuals with similar socio-economic/demographic characteristics) to the same class/cluster. Different clustering techniques can be used including heuristics, hierarchical, k-means, model-based clustering, etc. We opt to rely on two different clustering approaches, mixture model-based clustering and Gaussian Processes. In the former approach, each observation (e.g., socio-economic/demographic characteristics of a decision-maker) is assumed to be generated from a finite mixture of distributions where each distribution

represents a latent class/cluster (McLachlan et al., 2019). Most of the studies in Section 2.1.2 have used mixtures of distributions to represent the random distribution of taste parameters. Moreover, the majority of research on random heterogeneity has focused on improving the flexibility of utilities and parameter estimates. Instead, in this dissertation, we present an alternative mixture approach which consists of using a mixture of distributions to formulate the latent classes (rather than the choice model parameters) and improve their flexibility. In other words, instead of defining more complex distribution functions, we use, in the first proposed model (GBM-LCCM), a mixture of distributions to cluster decision-makers. In the machine learning community, this is known as mixture models and it is widely used as an unsupervised technique to cluster data into homogeneous groups/clusters. We aim to compare this approach to its traditional discrete choice model counterpart, the LCCM, in terms of its ability to estimate different classes and improve prediction accuracy while keeping model interpretability and being useful for policy testing and inferring economic indicators. The rationale for using model-based clustering in this dissertation, as opposed to other techniques, is threefold. First, a probabilistic method is needed to estimate the proposed latent class – choice model framework simultaneously as opposed to a two-stage sequential approach. The simultaneous estimation usually provides more efficient estimates than the sequential estimation. Second, mixture models allow more flexibility than the utility specification of latent classes which is usually defined to be linear in parameters. Third, such techniques provide a framework for evaluating the clusters, meaning that interpretability can be maintained to a large extent (Biernacki et al., 2000). However, the Gaussian-Bernoulli Mixture Model assumes that continuous and categorical variables are uncorrelated and as such the relations between the different

variables used for clustering cannot be well determined. Moreover, Gaussian-Bernoulli Mixture Models as well as simple neural networks are parametric models. Such models assume specific functional form for the distribution (or mapping function) and as such have predefined numbers of parameters that once learned (estimated) would be used for predictions. On one hand, parametric assumptions make the learning process (estimation) easier, faster and in less need of training data to learn the parameters. On the other hand, parametric models are constrained to the functional forms they assume, which restricts the flexibility of the learning process and might lead to poor generalization or prediction accuracies (Ghahramani, 2015).

In order to overcome the limitations of parametric clustering models while improving the flexibility and generalization performance of LCCM, nonparametric machine learning algorithms could be used instead. Such methods are data-driven, do not assume predefined functional forms and consequently are free to learn any functional form from the training data. These methods still contain parameters to control the complexity of the model rather than the functional form of the distribution (C. Bishop, 2006). Gaussian Process is one such method that avoids simple parametric assumptions and provides a fully Bayesian framework for modeling (Rasmussen & Williams, 2006). These characteristics make GPs very attractive for modeling uncertainties and complex nonlinear problems. Moreover, Gaussian Process can be mathematically equivalent to neural networks with very large number of hidden units (Neal, 1996). However, Gaussian Processes are generally easier to handle since the estimation of a neural network is usually complicated by the fact that the optimization problem might have several local optima while the posterior of the Gaussian Process for regression and classification is convex (Rasmussen & Williams, 2006). One

shortcoming of such models is that their nonparametric nature might make the LCCM less transparent at the class membership model. Note that, lately, Gaussian Processes have been receiving growing attention and are being applied to many regression and classification applications within transportation such as travel time prediction (Idé & Kato, 2009; Rodrigues et al., 2016), crowdsourced traffic data (Rodrigues et al., 2019; Rodrigues & Pereira, 2018), congestion and routing models (Liu et al., 2013), traffic volume forecasting (Xie et al., 2010), censored demand modeling (Gammelli et al., 2020), etc. However, in this dissertation, GPs are used as clustering techniques for mode choice modeling.

Finally, the two proposed models can be considered as new members of the Hybrid Choice Model (HCM) family since they deal with discrete latent variables/classes. The HCM was developed by several researchers (Ben-Akiva et al., 2002a, 2002b; McFadden, 1986; Walker and Ben-Akiva, 2002, to name a few) by incorporating latent variables in discrete choice models to model unobserved heterogeneity, improve goodness-of-fit and efficiency, extend policy relevance, and enhance behavioral realism (Abou-Zeid & Ben-Akiva, 2014). Moreover, the HCM is widely used to deal with the problems of combining Revealed Preferences (RP) data, Stated Preferences (SP) data, and psychometric indicators. This paper focuses on the LCCM family of the HCM's by defining the class membership model, first as a mixture model and second as a Gaussian Process. However, the problems of combining RP data with SP data and making use of attitudinal indicators to define latent variables are beyond the scope of this dissertation.

## CHAPTER 3

### GAUSSIAN-BERNOULLI MIXTURE LATENT CLASS CHOICE MODEL

In this chapter, we develop a hybrid model that consists of using Gaussian-Bernoulli Mixture Models (GBMMs), a model-based clustering approach, as a first-stage clustering tool to divide the population into homogenous groups/latent classes while utilizing discrete choice models to develop class-specific choice models.

Gaussian Mixture Models (GMMs) are widely used in machine learning, statistical analysis, pattern recognition, and data mining and can be easily formulated to define discrete latent variables (C. Bishop, 2006). GMM is a combination of  $K$  Gaussian densities where each density is a component (latent class) of the mixture and has its own mean vector and covariance structure. These models are more flexible than other clustering techniques (e.g., k-means or hierarchical clustering) since the covariance matrix of GMM can account for correlation between explanatory variables and clusters using different structures (McNicholas & Murphy, 2010). Particularly, the covariance matrices of GMM can have different structures such as: full covariance structure wherein each latent class has its own general covariance matrix, a diagonal covariance structure wherein each latent class has its own diagonal covariance matrix, a spherical structure wherein each latent class has its own single variance regardless of the number of explanatory variables, or a constrained version of one of the three previous structures (e.g., a tied covariance structure wherein all latent classes share the same general covariance matrix). We believe this flexible approach would help capture underlying behavioral heterogeneity and complex behavioral patterns within the population. However, GMM can only deal with continuous variables. Therefore, we

rely on a joint Gaussian-Bernoulli Mixture Model to assign decision-makers probabilistically to different latent classes using both continuous and discrete socio-economic characteristics while we make use of random utility models (e.g., logit models) for class-specific choice models. The full model is called Gaussian-Bernoulli Mixture - Latent Class Choice Models (GBM-LCCM). This is similar to the well-known LCCM that allows capturing heterogeneity in the choice process by allocating people to a set of  $K$  homogeneous classes.

The next Section (3.1) presents the LCCM formulation while the subsequent Section (3.2) develops the formulation and estimation technique of the proposed Gaussian-Bernoulli Mixture Latent Class Choice Model.

### 3.1. Latent Class Choice Model

LCCM consists of two components, a class membership model and a class-specific choice model. The class membership model estimates the probability that a decision-maker belongs to a specific class, typically as a function of his/her characteristics. The utility of belonging to latent class  $k$  for decision-maker  $n$  is defined as follows:

$$U_{nk} = S'_n \gamma_k + v_{nk} , \quad (6)$$

with  $S_n$  a vector of socio-economic/demographic variables of decision-maker  $n$  including a constant,  $\gamma_k$  the corresponding vector of unknown parameters that need to be estimated statistically using the available data, and  $v_{nk}$  a random disturbance term that is assumed to follow an independently and identically distributed (*iid*) Extreme Value Type I distribution over decision-makers and classes.



The probability of decision-maker  $n$  belonging to latent class  $k$  is then expressed as follows:

$$P(q_{nk} = 1 | S_n, \gamma_k) = \frac{e^{S_n' \gamma_k}}{\sum_{k'=1}^K e^{S_n' \gamma_{k'}}}, \quad (7)$$

with  $q_{nk}$  equal to 1 if decision-maker  $n$  belongs to latent class  $k$  and 0 otherwise.

Conditioned on the class membership of the decision-maker, the class-specific choice model formulates the probability of choosing a specific alternative as a function of the exogenous attributes of the alternatives. As such, the utility of decision-maker  $n$  choosing alternative  $j$  during time period / choice occasion  $t$ , conditional on him/her belonging to class  $k$ , is specified as:

$$U_{njt|k} = X_{njt}' \beta_k + \varepsilon_{njt|k}, \quad (8)$$

where  $X_{njt}$  is a vector of exogenous attributes related to alternative  $j$  during time period  $t$  and including a constant,  $\beta_k$  is the corresponding vector of unknown parameters that need to be estimated statistically using the available data, and  $\varepsilon_{njt|k}$  is a random disturbance term that is assumed to follow an *iid* Extreme Value Type I distribution over alternatives, decision-makers and classes.

Conditioned on class  $k$ , the probability of decision-maker  $n$  selecting an alternative  $j$  in time period  $t$  can then be written as follows:

$$P(y_{njt} = 1 | X_{njt}, q_{nk} = 1, \beta_k) = \frac{e^{V_{njt|k}}}{\sum_{j'=1}^J e^{V_{nj't|k}}}, \quad (9)$$

with  $J$  being the total number of alternatives.

Assuming that the conditional choice probabilities (Equation 9) for decision-maker  $n$  over all time periods  $T_n$  are conditionally independent, the conditional probability of observing a  $(J \times T_n)$  matrix of choices  $y_n$  can be expressed as follows:

$$P(y_n|X_n, q_{nk} = 1, \beta_k) = \prod_{t=1}^{T_n} \prod_{j=1}^J \left( P(y_{njt} = 1|X_{njt}, q_{nk} = 1, \beta_k) \right)^{y_{njt}}, \quad (10)$$

with  $X_n$  being a matrix consisting of  $J \times T_n$  vectors of  $X_{njt}$ ,  $y_n$  a  $(J \times T_n)$  matrix of all choices of individual  $n$  during all time periods  $T_n$  and consisting of choice indicators  $y_{njt}$ , and  $y_{njt}$  a choice indicator equal to 1 if decision-maker  $n$  chooses alternative  $j$  during time period  $t$  and 0 otherwise.

The unconditional probability of the observed choice of decision-maker  $n$  is then obtained by summing the product of the class membership probability (Equation 7) by the conditional choice probability (Equation 10) over all latent classes (we omit the dependencies on the left hand side of the equation to make the notation less cluttered):

$$P(y_n) = \sum_{k=1}^K P(q_{nk} = 1|S_n, \gamma_k) P(y_n|X_n, q_{nk} = 1, \beta_k). \quad (11)$$

Finally, the likelihood over a sample of independent decision-makers  $N$  is:

$$P(y) = \prod_{n=1}^N \sum_{k=1}^K P(q_{nk} = 1|S_n, \gamma_k) P(y_n|X_n, q_{nk} = 1, \beta_k). \quad (12)$$

### 3.2. Gaussian-Bernoulli Mixture Latent Class Choice Model

We propose to replace the class membership model,  $P(q_{nk} = 1|S_n, \gamma_k)$ , by a Gaussian-Bernoulli Mixture Model (GBM), a probabilistic machine learning approach used for clustering (Figures 2 and 3) where a Gaussian Mixture Model (GMM) is used for continuous variables and a Bernoulli Mixture Model (BMM) for discrete/binary variables. We split the vector of characteristics of decision-maker  $n$  ( $S_n$ ) into two sub-vectors,  $S_{cn}$  and  $S_{dn}$ .  $S_{cn}$  accounts for the continuous characteristics of decision-maker  $n$  with dimension  $D_c$  equal to the number of elements in  $S_{cn}$  while  $S_{dn}$  accounts for the

discrete/binary characteristics of decision-maker  $n$  with dimension  $D_d$  equal to the number of elements in  $S_{dn}$ .

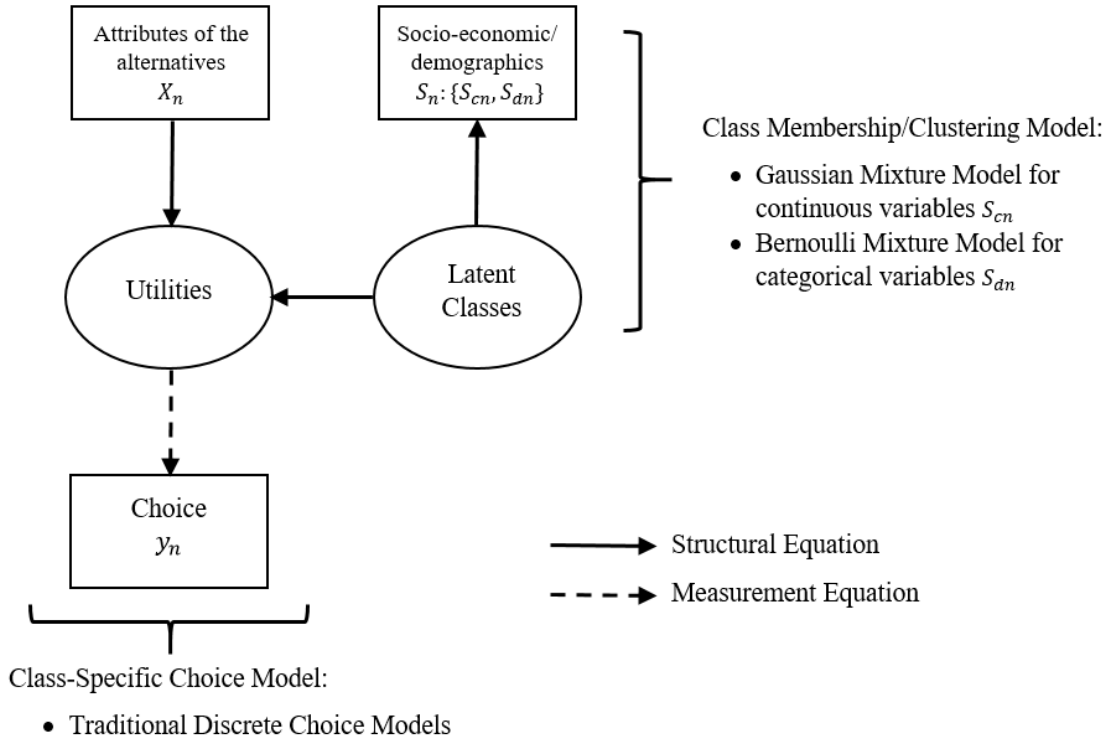


Figure 2: Gaussian-Bernoulli Mixture - Latent Class Choice Model (GBM-LCCM)

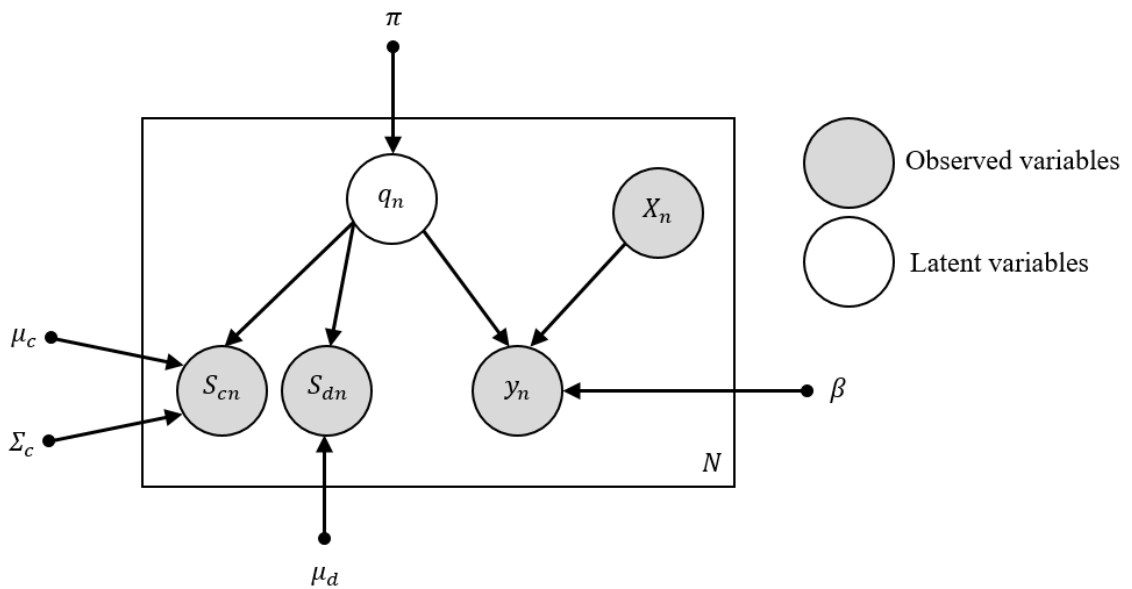


Figure 3: Graphical Representation of the proposed Gaussian-Bernoulli Mixture Latent Class Choice Model (GBM-LCCM) for a set of  $N$  decision-makers

GMM is a combination of  $K$  Gaussian densities where each density,  $\mathcal{N}(S_{cn}|\mu_{ck}, \Sigma_{ck})$ , is a component of the mixture and has its own mean  $\mu_{ck}$  (with dimension equal to the number of elements in  $S_{cn}$ ), covariance  $\Sigma_{ck}$ , and mixing coefficient  $\pi_k$  (the overall probability that an observation comes from component  $k$ ) (C. Bishop, 2006). BMM is a combination of  $K$  mixture components where each component  $k$  is a product of  $D_d$  independent Bernoulli probability functions and has its own mean vector  $\mu_{dk}$ .

Replacing the class membership probability by a GBM is not a straightforward task. The probability of decision-maker  $n$  belonging to class  $k$ ,  $P(q_{nk} = 1|S_n)$ , is the posterior probability of the GBM and cannot be part of the likelihood function that needs to be maximized. Instead, we estimate the probability of observing decision-maker  $n$  with characteristics  $S_n = \{S_{cn}, S_{dn}\}$  given that he/she belongs to latent class  $k$  (Figure 2). Note that in Figure 2 the causality goes from the latent classes to the socio-economic variables and not the other way around as in the traditional LCCM (Figure 1). This stems from the fact that Gaussian-Bernoulli Mixture Models are generative models that learn the joint probability of the features/characteristics ( $S_n$ ) and the labels/classes ( $q_{nk}$ ) then make use of Bayes' theorem to calculate the posterior probability  $P(q_{nk} = 1|S_n)$  (C. Bishop, 2006). We follow the same steps to estimate the proposed GBM-LCCM. First, we estimate the joint probability of the model then we calculate the posterior and marginal probabilities by using Bayes' rules. The graphical representation of the proposed model is shown in Figure 3.

Given the conditional independence properties of the graphical model structure of the proposed model and assuming that the continuous and binary data of the Gaussian and Bernoulli distributions are independent, the joint probability of  $S_{cn}, S_{dn}$ ,

$y_n$  and  $q_{nk}$  can be specified as the product of the class probability (first term on the right hand side below), the densities of  $S_{cn}$  and  $S_{dn}$  conditional on the class (second and third terms) and the choice probability conditional on the class (fourth term), as follows:

$$\begin{aligned} P(S_{cn}, S_{dn}, y_n, q_{nk} = 1 | X_n, \beta_k, \pi_k, \mu_{ck}, \Sigma_{ck}, \mu_{dk}) \\ = P(q_{nk} = 1 | \pi_k) P(S_{cn} | q_{nk} = 1, \mu_{ck}, \Sigma_{ck}) P(S_{dn} | q_{nk} = 1, \mu_{dk}) \quad (13) \\ \times P(y_n | X_n, q_{nk} = 1, \beta_k), \end{aligned}$$

where:

$$P(q_{nk} = 1 | \pi_k) = \pi_k, \quad (14)$$

$$\sum_{k=1}^K \pi_k = 1, \quad (15)$$

$$\begin{aligned} P(S_{cn} | q_{nk} = 1, \mu_{ck}, \Sigma_{ck}) &= \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \\ &= \frac{1}{\sqrt{(2\pi)^{D_c} |\Sigma_{ck}|}} \exp\left(-\frac{1}{2} (S_{cn} - \mu_{ck})' \Sigma_{ck}^{-1} (S_{cn} - \mu_{ck})\right), \quad (16) \end{aligned}$$

$$P(S_{dn} | q_{nk} = 1, \mu_{dk}) = \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})}, \quad (17)$$

with  $|\Sigma_{ck}|$  the determinant of the covariance matrix,  $S_{dni}$  a binary characteristic of decision-maker  $n$  and  $\mu_{dk_i}$  its corresponding mean.

The joint probability of  $S_{cn}$ ,  $S_{dn}$  and  $y_n$  can be then obtained by taking the marginal of Equation (13) over all components  $K$ :

$$\begin{aligned} P(S_{cn}, S_{dn}, y_n | X_n, \beta, \pi, \mu_c, \Sigma_c, \mu_d) \\ = \sum_{k=1}^K P(S_{cn}, S_{dn}, y_n, q_{nk} = 1 | X_n, \beta_k, \pi_k, \mu_{ck}, \Sigma_{ck}, \mu_{dk}), \quad (18) \end{aligned}$$

where  $\beta$  is a matrix consisting of  $K$  vectors of  $\beta_k$ ,  $\pi$  is a vector consisting of  $K$  mixing coefficients  $\pi_k$ ,  $\mu_c$  is a matrix consisting of  $K$  mean vectors of continuous variables  $\mu_{ck}$ ,

$\Sigma_c$  is a structure consisting of  $K$  covariance matrices  $\Sigma_{ck}$ , and  $\mu_d$  is a matrix consisting of  $K$  mean vectors of discrete variables  $\mu_{dk}$ .

Finally, the likelihood function of the proposed hybrid model for all decision-makers  $N$  is formulated as follows (we omit the dependencies on the left hand side of the equation to make the notation less cluttered):

$$\begin{aligned}
P(S_c, S_d, y) &= \prod_{n=1}^N P(S_{cn}, S_{dn}, y_n | X_n, \beta, \pi, \mu_c, \Sigma_c, \mu_d) \\
&= \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})} \\
&\quad \times \prod_{t=1}^{T_n} \prod_{j=1}^J \left( \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_k}} \right)^{y_{njt}}.
\end{aligned} \tag{19}$$

Usually, traditional discrete choice models are estimated using maximum likelihood estimation techniques which aim at maximizing the likelihood of the observed data given the model parameters. However, maximizing the log-likelihood of both LCMM and GBM-LCCM is a complex task due to the summation over  $k$  that will appear inside the logarithm of Equations 12 and 19. Setting the derivatives of the log-likelihood to zero will not lead to a closed-form solution (C. Bishop, 2006). Moreover, maximizing the log-likelihood of a discrete choice model with discrete latent variables, such as LCCM, becomes more difficult as the number of classes, and consequently the number of parameters, increases. With larger number of classes and parameters, the calculation of the gradient becomes slower and empirical singularity might arise at some iterations making the inversion of the Hessian matrix numerically challenging (Train, 2008).

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a two-stage iterative maximization technique that overcomes the aforementioned complications by repeatedly maximizing an expectation or a lower bound function of the likelihood (Train, 2008). Several studies have applied EM algorithms to discrete choice model applications with mixing distributions or discrete latent variables (Bhat, 1997; El Zarwi, 2017b; Train, 2008, to name a few) and the results showed that such techniques are computationally attractive. The EM algorithm framework consists of two main steps, E (expectation) and M (maximization). In the former (E-step), the expectations of the latent variables conditioned on the current estimates of the unknown parameters and the observed variables are estimated, while in the latter (M-step), the expectation of the log-likelihood is maximized, conditioned on the observed variables and the expectations of the latent variables obtained from the E-step, to update the estimates of the unknown parameters. The algorithm alternates between these two steps until a predefined convergence criterion is satisfied. In this dissertation, an EM-based algorithm is derived and implemented for estimating the proposed GBM-LCCM.

### 3.2.1. EM Algorithm

The first step of the EM algorithm requires writing the joint likelihood function (Equation 19) assuming that the clusters (latent classes,  $q_{nk}$ ) are observed:

$$\begin{aligned}
 P(S_c, S_d, y, q) = & \prod_{n=1}^N \prod_{k=1}^K \left[ \pi_k \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})} \right]^{q_{nk}} \\
 & \times \prod_{n=1}^N \prod_{k=1}^K \prod_{t=1}^{T_n} \prod_{j=1}^J \left[ \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_k}} \right]^{y_{njt} q_{nk}} .
 \end{aligned} \tag{20}$$

Taking the logarithm of the likelihood, the function breaks into two separate parts, one for each of the two sub-models (class membership model and class-specific choice model), as follows:

$$\begin{aligned}
LL = & \sum_{n=1}^N \sum_{k=1}^K q_{nk} \log \left[ \pi_k \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})} \right] \\
& + \sum_{n=1}^N \sum_{k=1}^K \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} q_{nk} \log \left[ \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_k}} \right].
\end{aligned} \tag{21}$$

Now, the unknown parameters  $\{\mu_{ck}, \Sigma_k, \mu_{dk}, \pi_k, \beta_k\}$  of each component  $k$  can be found by setting the derivatives of the above log-likelihood with respect to each of the unknown parameters to zero if and only if  $q_{nk}$  is known. To find the values of  $q_{nk}$ , we estimate the expectation of  $q_{nk}$  (E-step) using Bayes' theorem:

$$\begin{aligned}
P(q_{nk} = 1 | y_n, S_{cn}, S_{dn}, X_n, \mu_{ck}, \Sigma_{ck}, \mu_{dk}, \pi_k, \beta_k) \\
& \propto P(q_{nk} = 1 | \pi_k) P(S_{cn} | q_{nk} = 1, \mu_{ck}, \Sigma_{ck}) P(S_{dn} | q_{nk} = 1, \mu_{dk}) \\
& \quad \times P(y_n | X_n, q_{nk} = 1, \beta_k) \\
& \propto \pi_k \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})} \\
& \quad \times \prod_{t=1}^{T_n} \prod_{j=1}^J \left[ \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_k}} \right]^{y_{njt}},
\end{aligned} \tag{22}$$

$$E[q_{nk}] = \gamma_{q_{nk}}$$

$$\begin{aligned}
& \frac{\pi_k \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \prod_{i=1}^{D_d} \mu_{dk_i}^{S_{dni}} (1 - \mu_{dk_i})^{(1-S_{dni})} \prod_{t=1}^{T_n} \prod_{j=1}^J \left[ \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_k}} \right]^{y_{njt}}}{\sum_{k'=1}^K \left[ \pi_{k'} \mathcal{N}(S_{cn} | \mu_{ck'}, \Sigma_{ck'}) \prod_{i=1}^{D_d} \mu_{dk'_i}^{S_{dni}} (1 - \mu_{dk'_i})^{(1-S_{dni})} \prod_{t=1}^{T_n} \prod_{j=1}^J \left[ \frac{e^{X'_{njt} \beta_{k'}}}{\sum_{j'=1}^J e^{X'_{nj't} \beta_{k'}} \right]^{y_{njt}} \right]}
\end{aligned} \tag{23}$$



It is to be noted that  $\pi_k$  (Equation 14) is considered as the prior probability of  $q_{nk} = 1$  while  $\gamma_{q_{nk}}$  (Equation 23) is the corresponding posterior probability.

Next, the likelihood should be maximized to find the unknown parameters. However, since Equation 21 cannot be maximized directly due to the presence of latent variables  $q_{nk}$ , we consider instead the expected value of the log-likelihood, where the expectation is taken w.r.t.  $q_{nk}$ .

Making use of Equations 21 and 23, gives:

$$\begin{aligned}
E[LL] = & \sum_{n=1}^N \sum_{k=1}^K \gamma_{q_{nk}} \left( \log \pi_k + \log \mathcal{N}(S_{cn} | \mu_{ck}, \Sigma_{ck}) \right. \\
& \left. + \sum_{i=1}^{D_d} [S_{dni} \log \mu_{dki} + (1 - S_{dni}) \log(1 - \mu_{dki})] \right) \\
& + \sum_{n=1}^N \sum_{k=1}^K \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} \gamma_{q_{nk}} \log \left[ \frac{e^{x'_{njt} \beta_k}}{\sum_{j'=1}^J e^{x'_{nj't} \beta_k}} \right].
\end{aligned} \tag{24}$$

Setting the derivatives of the expected log-likelihood with respect to the unknown parameters to zero, we obtain the solutions of the unknown parameters as follows:

$$\mu_{ck} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{q_{nk}} S_{cn}, \tag{25}$$

$$\Sigma_{ck} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{q_{nk}} (S_{cn} - \mu_{ck})(S_{cn} - \mu_{ck})', \tag{26}$$

$$\mu_{dk} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{q_{nk}} S_{dn}, \tag{27}$$

$$\pi_k = \frac{N_k}{N}, \tag{28}$$

$$\beta_k = \operatorname{argmax}_{\beta_k} \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} \gamma_{q_{nk}} \log \left[ \frac{e^{x'_{njt} \beta_k}}{\sum_{j'=1}^J e^{x'_{njt} \beta_k}} \right], \quad (29)$$

where we have defined:

$$N_k = \sum_{n=1}^N \gamma_{q_{nk}}. \quad (30)$$

Equations 25 to 28 are the closed-form solutions of the Gaussian mean matrix, Gaussian covariance matrix, Bernoulli mean matrix, and mixing coefficients, respectively. As for the parameters  $\beta_k$  (Equation 29), no closed-form solution can be obtained. Instead, we resort to the gradient-based numerical optimization method BFGS (Nocedal et al., 1999).

To sum up, the EM algorithm alternates between the E-step and M-step until convergence is reached. First, we initialize the unknown parameters. Second, we estimate the expected values of the latent variables using Equation 23 (E-step). Next, we update the values of the unknown parameters using Equations 25 to 29 (M-step). Finally, we evaluate the log-likelihood using the current values of the unknown parameters and check for convergence. If the convergence criterion is not met, we return to the E-step.

### 3.2.2. Final Likelihood

After reaching convergence, we evaluate the marginal probability  $P(y)$  of observing a vector of choices  $y$  of all decision-makers  $N$  as follows (we omit the dependencies on the left hand side of the equation to make the notation less cluttered):

$$\begin{aligned}
P(\mathbf{y}) &= \prod_{n=1}^N \sum_{k=1}^K P(q_{nk} = 1 | S_{cn}, S_{dn}, \mu_{ck}, \Sigma_{ck}, \mu_{dk}, \pi_k) \\
&\quad \times \prod_{t=1}^{T_n} \prod_{j=1}^J \left( P(y_{njt} = 1 | X_{njt}, q_{nk} = 1, \beta_k) \right)^{y_{njt}},
\end{aligned} \tag{31}$$

where  $P(q_{nk} = 1 | S_{cn}, S_{dn}, \mu_{ck}, \Sigma_{ck}, \mu_{dk}, \pi_k)$  is the posterior probability of vector  $S_n = \{S_{cn}, S_{dn}\}$  being generated by cluster  $k$ . The posterior probability can be formulated using Bayes' theorem:

$$\begin{aligned}
&P(q_{nk} = 1 | S_{cn}, S_{dn}, \mu_{ck}, \Sigma_{ck}, \mu_{dk}, \pi_k) \\
&= \frac{P(q_{nk} = 1 | \pi_k) P(S_{cn} | q_{nk} = 1, \mu_{ck}, \Sigma_{ck}) P(S_{dn} | q_{nk} = 1, \mu_{dk})}{\sum_{k'=1}^K P(q_{nk'} = 1 | \pi_{k'}) P(S_{cn} | q_{nk'} = 1, \mu_{ck'}, \Sigma_{ck'}) P(S_{dn} | q_{nk'} = 1, \mu_{dk'})}.
\end{aligned} \tag{32}$$

The above marginal probability (Equation 31) is used for comparing the GBM-LCCM with the traditional LCCM (Equation 12) and for calculating out-of-sample prediction accuracies.

Note that, in case only continuous socio-economic characteristics are used, the proposed model becomes Gaussian Mixture - Latent Class Choice Model (GM-LCCM). The formulation would follow the same steps of section 3.2 but without the mixture of Bernoulli distribution functions (Equation 17). The same applies in case only discrete variables are used in the clustering stage.

# CHAPTER 4

## GBM-LCCM APPLICATIONS

In this chapter, we develop and present two applications of the proposed modeling approach (GBM-LCCM) using two different case studies on travel mode choice behavior. The chapter is organized as follows. Section 4.1 describes the two datasets that are used to compare the GBM-LCCM with different benchmark models. Section 4.2 discusses the implementation of the different models. Section 4.3 presents their formulation/specification and estimation results. Section 4.4 concludes.

### **4.1. Data**

Two different datasets on travel mode choice behavior are used in this chapter in order to assess the proposed model. The first one is a Revealed Preferences (RP) dataset of individual daily trips (Section 4.1.1) while the second one is a Stated Preferences (SP) dataset of weekly travel choices (Section 4.1.2).

#### ***4.1.1. London Dataset***

The first application is based on the “London” dataset which is available online as supplementary material to Hillel et al. (2018). The dataset combines individuals’ trip diaries of the London Travel Demand Survey (LTDS) from April 2012 to March 2015 with their corresponding modes alternatives extracted from a Google directions application programming interface (API) and corresponding estimates of car operating costs and public transport fares. The dataset consists of 81,086 trips, four modes (walking, cycling, public transport, and driving), and different trip purposes (e.g.,

Home-Based Work, Home-Based Education, etc.). In this chapter, we only consider Home-Based Work (HBW) trips and trips made by car and public transport in order to have a balanced sample. The first two years (7,814 trips) are used for estimation/training while the third year (3,883 trips) is used for testing/prediction.

#### **4.1.2. AUB Dataset**

The second application is based on a dataset from the American University of Beirut (AUB) in Lebanon, a major private university with about 8,094 students, 4,173 staff, and 2,168 faculty members (“AUB Fact Book 2016-2017,” 2016). The university is located in a dense urban area within Municipal Beirut and its surrounding neighborhood suffers from high levels of congestion and parking demand. To overcome these problems, AUB was considering a few years ago two alternative sustainable transport modes for its population, shared-taxi and shuttle services. The shared-taxi would be a door-to-door service that provides on-demand transport between AUB gates and users’ residences (and vice versa) while the shuttle service would be a non-stop first/last mile service between AUB gates and satellite parking hubs (and vice versa) where commuters could park their cars just a few kilometers away from AUB. In order to investigate the willingness of the AUB population to use the new transport services if they were implemented, a web-based stated preferences commuting survey was designed and sent to all AUB students, faculty members, and staff in April of 2017. The survey collected information about each respondent’s daily travel to and from AUB, place of residence, and socio-economic characteristics. In addition, the stated preferences survey offered each respondent four hypothetical scenarios in which he/she had to state how many weekdays per week he/she is willing to use the two proposed

services in addition to his/her current mode of commute. An example of the hypothetical scenarios is shown in Figure 4. A sub-sample of car users who come five days per week to AUB is used in this application. The sub-sample consists of 650 respondents and 2,600 choice observations. For more details about the dataset and the survey design, readers may refer to Sfeir et al. (2020).

Shared-Taxi	Shuttle	Your Current Commute to AUB
<u>Door-to-door travel time</u> <b>33 min</b> <u>Waiting time for late pick-up and/or early drop-off</u> <b>0 to 5 min</b> <u>Number of passengers sharing a ride</u> <b>4 to 6 (Minivan)</b>	<u>Access travel time</u> <b>12 min (by car)</b> <u>Frequency</u> <b>Every 5 min</b> <u>In-Shuttle travel time</u> <b>27 min</b>	<u>Total Travel Time</u> <b>30 min</b> <u>Mode of Travel</u> <b>Car</b>
<u>Mobile App/Wi-Fi/Live tracking</u> <b>Available</b>	<u>Wi-Fi/Live tracking</u> <b>Not Available</b>	
<u>One-way fare</u> <b>4,000 L.L.</b>	<u>One-way shuttle fare including parking cost</u> <b>1,000 L.L.</b> <u>One-way fuel cost (for access by car)</u> <b>700 L.L.</b>	<u>Parking cost</u> <b>5,000 L.L.</b> <u>One-way fuel cost</u> <b>1,800 L.L.</b>

Based on this scenario, and considering your current pattern to AUB, how many weekdays per week will you use the proposed services? Remember that you indicated earlier that you come on 5 weekdays per week to AUB.

Shared-Taxi	0, 1, 2, 3, 4, 5
Shuttle	0, 1, 2, 3, 4, 5
Your Current Commute to AUB	0, 1, 2, 3, 4, 5

Figure 4: Hypothetical scenario and choice question example from the survey

## 4.2. Implementation

In addition to the proposed model, we estimate and present LCCM models to benchmark the proposed GBM-LCCM against its traditional discrete choice model counterpart. The proposed model (GBM-LCCM) is implemented in Python by using some blocks from: 1) the lccm package (El Zarwi, 2017a, 2017b), a python package that

implements an EM algorithm for estimating traditional Latent Class Choice Models; 2) and the Gaussian Mixture class of the Scikit-Learn library (Pedregosa et al., 2011). The code is publicly available on GitHub<sup>2</sup>. The traditional LCCMs are also estimated in Python using the lccm package (El Zarwi, 2017a, 2017b). Convergence of both EM implementations, the proposed model and the traditional LCCM, is assumed to be reached once the change in the log-likelihood function is smaller than  $1 \times 10^{-4}$ .

The EM algorithm has proved to be a powerful approach for estimating models with latent variables or missing data (Bhat, 1997; Train, 2008). However, the algorithm is sensitive to starting values and might not guarantee convergence to the global maximum. Therefore, a good set of initial values is of great importance to assure proper convergence. Different approaches and heuristics have been used in the literature to overcome this limitation. The two most used approaches are random initialization and incremental initialization, where estimates of models with  $K - 1$  classes are used as starting values for models with  $K$  classes. In this dissertation, we make use of both approaches (Table 2). In addition, the Gaussian-Bernoulli Mixture Models are initialized randomly and using  $k$ -means, a deterministic unsupervised machine learning approach. In total, each model is estimated 25 times and the log-likelihood variance of the different runs is reported to check if the model is converging to the same solution or not.

---

<sup>2</sup> <https://github.com/gsfair/GBM-LCCM>

Table 2: EM Initialization

<b>GBM-LCCM</b>		
<b>Class Membership Model</b>	<b>Class Specific Choice Model</b>	<b>Trials</b>
Random	0	5
Random	Random	5
K-means	0	5
K-means	Random	5
Random/K-means	Estimates of K-1 model and 0/Random for the additional class	5
<b>LCCM</b>		
<b>Class Membership Model</b>	<b>Class Specific Choice Model</b>	<b>Trials</b>
0	0	5
0	Random	5
Random	0	5
Random	Random	5
Estimates of K-1 model and 0/Random for the additional class		5

### 4.3. Applications

We present the specification of the models and estimation results of the two datasets in Sections 4.3.1 and 4.3.2, respectively.

#### 4.3.1. London Case Study

Using the first dataset, we test and present three different trials and, for the sake of brevity, we only present in this section summary statistics of the estimated models. Details of all the estimated models are presented in Appendix A. In each trial, we compare the new approach to the traditional LCCM with the same specification in order to ensure that any potential differences in results are not attributed to model specification.



#### 4.3.1.1. First Trial

##### Model Specification

We assume that the latent classes of the GBM-LCCM are characterized by the available socio-economic variables age, gender, car ownership, and driving license, such that  $age_n$  is a continuous variable representing the age of decision-maker  $n$ ;  $female_n$  is a binary variable that equals to 1 if decision-maker  $n$  is female and 0 otherwise;  $car\_own_{n1}$  a binary variable that equals to 1 if the number of cars in the household of decision-maker  $n$  is more than 0 but less than one per adult and 0 otherwise,  $car\_own_{n2}$  equals to 1 if the number of cars in the household of decision-maker  $n$  is one or more per adult and 0 otherwise; and  $license_n$  is a binary variable that equals to 1 if decision-maker  $n$  has a driving license and 0 otherwise. Since only one continuous variable (age) is used for clustering, two covariance structures, full and tied, are tested. Regarding the class-specific choice models, we only consider alternative-specific travel time and travel cost coefficients in addition to a constant in the utility of the car alternative.

##### Results

Table 3 presents summary statistics of the new approach and the traditional LCCM<sup>3</sup>. We show the average joint log-likelihood of the GBM-LCCM, the average marginal (i.e., choice) log-likelihood of the two models, the corresponding Akaike Information Criterion (AIC)<sup>4</sup> and Bayesian Information Criterion (BIC)<sup>5</sup>, the predictive log-likelihood (for the test sample), and the variance of the marginal log-likelihood of

---

<sup>3</sup> Several MNL and Mixed Logit models were also estimated for this application but showed no clear advantage over the GBM-LCCM in terms of predictive power and as such are not shown in this dissertation.

<sup>4</sup>  $AIC = 2M - 2 \log LL$ , where  $M$  is the number of parameters and  $LL$  is the marginal choice log-likelihood.

<sup>5</sup>  $BIC = M \log D - 2 \log LL$ , where  $D$  is the number of data points (observations).

the LCCM and GBM-LCCM to evaluate the stability of the EM solutions since these models are run multiple times with different starting values.

We first look at the marginal log-likelihoods of all the estimated models. The LCCM ( $K = 3$ ) has the best log-likelihood (-2,470.01). However, the log-likelihood variance of the LCCM with three latent classes is very high (190.64) meaning the model did not converge to the same solution and should be neglected. Moreover, the two LCCMs ( $K = 2$  and  $K = 3$ ) have positive public transport cost coefficients. Such models should be ignored since models with counter-intuitive coefficient signs cannot be used for meaningful predictions and policy analysis. For the LCCM approach, no more than three classes are estimated mainly due to identification problems (very high standard errors for the class-specific parameter estimates). Regarding the proposed GBM-LCCM, the full covariance model with 5 latent classes has also a high variance and should be ignored. After eliminating all unstable models (i.e., high variance) and models with unexpected coefficient signs, the two GBM-LCCM with a tied and full covariance structure, respectively, and four latent classes can be selected as the best models since they have the best LL, AIC, BIC, and predictive power. However, it is to be noted that models with tied and full covariance structures have similar performance, with slightly better log-likelihood for the model with a full covariance structure and slightly better predictive power for the model with a tied structure. This is mainly due to the additional parameters from the full covariance structure of the GMM.

Next, the latent classes of the tied-GBM-LCCM with four classes are described based on the mean matrix of the Gaussian-Bernoulli Mixture Model (Table 4). Note that the continuous variable age is standardized. Therefore, a negative value means the latent

class is characterized by young individuals while a positive value means individuals are older than the average (which is 40 years).

*K1: Licensed drivers in their forties*

The first latent class is characterized by individuals with an age near the population average (40 years) since the mean of age is around 0. Individuals belonging to this class are mostly licensed drivers (91.9%) and living in households with low car ownership.

*K2: Young with low car ownership*

The second latent class has the youngest individuals ( $\mu_{age} < 0$ ) from both genders who are almost equally likely to be licensed (48.8%) or unlicensed drivers (51.2%). In addition, individuals belonging to this class live in households with no cars (55.1%) or less than one car per adult (42.4%).

*K3: Licensed elderly*

This class includes the oldest individuals (highest  $\mu_{age}$  across all classes) who are mostly males (72.1%), licensed drivers (94.5%), and belong to families with less than one car per adult (98.6%).

*K4: Licensed drivers with high car ownership*

The last latent class is characterized by old individuals from both genders. Moreover, individuals are licensed drivers (99%) who live in households with more than one car per adult (98.2%).

The above analysis is a strong indication that the proposed model provides a simple interpretability at the class membership level, although the random utility formulation of the latent classes is replaced by a full mixture model.

Finally, Table 5 presents the class-specific parameter estimates of the tied-GBM-LCCM model with four classes. All cost and travel time parameters have the expected negative sign. Individuals from the first, third, and fourth classes are insensitive to travel cost of public transport.

#### 4.3.1.2. Second Trial

##### Model Specification

In the second trial, we adopt the same class membership specification as in the first trial but a more complex class-specific choice utilities' specification. In particular, public transport travel time is included in the utilities as three separate attributes: access travel time (walking time between origin and first public transport stage, and final public transport stage and destination), bus/rail travel time (travel time spent on rail and bus services), and interchange travel time (walking and waiting time at the stop for interchanges on public transport route). However, all five models generated positive public transport cost coefficients. A logarithmic specification of public transport cost was used in the class-specific choice utilities in order to resolve the issue of counter-intuitive sign of cost coefficients.

##### Results

Table 6 presents summary statistics of the new approach and its LCCM counterpart. The logarithmic transformation of public transport cost did solve the issue of counter-intuitive signs for the GBM-LCCM with two and three latent classes but it had no impact on the remaining models. In addition, the LCCM and GBM-LCCM models with positive cost coefficients showed very high convergence instability (very high log-likelihood variances). After eliminating all models with counter-intuitive coefficient signs and high log-likelihood variances, we can select the tied-GBM-LCCM

with three latent classes as the best model since it has the best AIC, BIC, and predictive log-likelihood. Note that the results of this specification are consistent with that of section 4.3.1.1. The tied and full covariance structure models have similar performance with slight differences in goodness-of-fit and predictive measures due to the differences in the covariance structures of the GMM.

#### 4.3.1.3. Third trial

##### Model Specification

For the third and last attempt, we consider the same latent classes' formulation and class-specific choice utilities' specification as in the first trial (section 4.3.1.1). In addition, we include in the class-specific choice utilities four additional variables (start time, day of week, month, and traffic variability). Traffic variability is added to the utilities as a continuous variable while the remaining three variables are binned and included as dummy variables. We use the same bins that are defined by Hillel et al. (2019). The start time of the trips is grouped into four categories: AM peak (06:30-09:29), inter-peak (09:30-16:29), peak (16:30-19:29), and night (19:30-06:29). The day of the week is divided into week days (Monday to Friday), Saturday, and Sunday. Finally, the trip month is grouped into winter season (December to February) and all other months (March to November).

##### Results

Table 7 presents summary statistics of all estimated models. The LCCM ran into identification issues (class-specific choice parameter estimates with very large standard errors) while the GBM-LCCM was able to determine only two latent classes. The Gaussian-Bernoulli mixture formulation of the latent classes showed a superior

clustering ability by determining two homogeneous groups within the sample while the traditional random utility formulation of the LCCM had computational problems.

Table 3: First trial

	<b>K</b>	<b>Joint LL<sup>a</sup></b>	<b>LL<sup>b</sup></b>	<b>Variance<sup>c</sup></b>	<b>AIC</b>	<b>BIC</b>	<b>Pred. LL</b>	<b>Notes</b>
LCCM	2		-2,643.15	0	5,318.30	5,429.72	-1,234.47	$\beta_{cost\_pt2} = 0.0618$ (p = 0.14)
	3		-2,470.01	190.64	4,994.02	5,182.04	-1,182.95	$\beta_{cost\_pt2} = 0.0748$ (p = 0.57)
GBM-LCCM Full Covariance	2	-18,660.22	-2,920.92	0	5,887.84	6,048.00	-1,387.89	
	3	-17,502.93	-2,807.87	0	5,685.74	5,929.47	-1,300.33	
	4	<b>-17,390.22</b>	<b>-2,703.27</b>	0	5,500.54	5,827.83	-1,262.86	
	5	-17,148.88	-2,671.04	27.45	5,460.08	5,870.94	-1,266.22	$\beta_{cost\_pt1} = 0.0258$ (p = 0.86)
GBM-LCCM Tied Covariance	2	-18,662.68	-2,920.81	0	5,885.62	6,038.82	-1,387.40	
	3	-17,505.89	-2,807.91	0	5,681.82	5,911.62	-1,300.17	
	4	-17,393.78	-2,703.38	<b>0</b>	<b>5,494.76</b>	<b>5,801.16</b>	<b>-1,260.08</b>	

a: joint log-likelihood of the GBM-LCCM (Equation 21)

b: marginal log-likelihood of the GBM-LCCM (Equation 31) and the LCCM (Equation 12)

c: Marginal log-likelihood (LL) variance of the different runs

Table 4: Mean matrix of the class membership model (GBM) – Tied Covariance – K = 4

Parameter		Class 1	Class 2	Class 3	Class 4
<b>age</b> **	Continuous	0.034	-0.285	0.447	0.332
<b>female</b>	Yes	0.426	0.533	0.279	0.463
	No*	0.574	0.467	0.721	0.537
<b>license</b>	Yes	0.919	0.488	0.945	0.990
	No*	0.081	0.512	0.055	0.010
<b>car_own<sub>0</sub></b>	0*	0.056	0.551	0.014	0.018
<b>car_own<sub>1</sub></b>	] 0 – 1 [	0.944	0.424	0.986	0
<b>car_own<sub>2</sub></b>	≥ 1	0	0.025	0	0.982

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Table 5: Parameter estimates of the class-specific choice models – Tied Covariance – K = 4

Parameter	Class 1	Class 2	Class 3	Class 4
<b>ASC (Car)</b>	2.35 (0.00)	-0.858 (0.00)	2.51 (0.00)	1.52 (0.00)
<b>Travel Time (PT)</b>	-0.178 (0.00)	-0.0751 (0.00)	-0.112 (0.00)	-0.0646 (0.00)
<b>Travel Time (Car)</b>	-0.316 (0.00)	-0.284 (0.00)	-0.115 (0.00)	-0.106 (0.00)
<b>Cost (PT)</b>	-0.102 (0.28)	-0.267 (0.01)	-0.106 (0.49)	-0.0206 (0.63)
<b>Cost (Car)</b>	-0.492 (0.00)	-0.181 (0.06)	-0.207 (0.00)	-0.153 (0.00)

Values within parentheses are p-values

Travel Time variables are in minutes

Cost variables are in Pound Sterling (£ gbp)



Table 6: Second trial

	<b>K</b>	<b>Joint LL<sup>a</sup></b>	<b>LL<sup>b</sup></b>	<b>Variance<sup>c</sup></b>	<b>AIC</b>	<b>BIC</b>	<b>Pred. LL</b>	<b>Notes</b>
<b>LCCM</b>	2		-2,633.56	880.40	5,307.12	5,446.39	-1,223.45	$\beta_{Log\_cost\_pt2} = 0.0134$ (p = 0.55)
	3		-2,458.11	750.66	4,982.22	5,212.02	-1,192.74	$\beta_{Log\_cost\_pt1} = 0.188$ (p = 0.00)
<b>GBM-LCCM Full Covariance</b>	2	-18,646.92	-2,906.82	0	5,867.64	6,055.66	-1,391.48	
	3	<b>-17,485.77</b>	<b>-2,790.88</b>	0	5,663.76	5,949.27	-1,302.31	
	4	-17,365.19	-2,684.59	2,712.20	5,479.18	5,862.18	-1,271.95	$\beta_{Log\_cost\_pt1} = 0.0779$ (p = 0.33)
<b>GBM-LCCM Tied Covariance</b>	2	-18,650.17	-2,907.00	0	5,866.00	6,047.06	-1,390.03	
	3	-17,489.31	-2,791.21	0	<b>5,660.42</b>	<b>5,932.00</b>	<b>-1,301.78</b>	
	4	-17,372.80	-2,684.57	95.95	5,473.14	5,835.25	-1,260.96	$\beta_{Log\_cost\_pt3} = 0.0234$ (p = 0.82)

a: joint log-likelihood of the GBMLCCM model (Equation 21)

b: marginal log-likelihood of the GBMLCCM (Equation 31) and the LCCM (Equation 12)

c: Marginal log-likelihood (LL) variance of the different runs

Table 7: Third trial

	<b>K</b>	<b>Joint LL<sup>a</sup></b>	<b>LL<sup>b</sup></b>	<b>Variance<sup>c</sup></b>	<b>AIC</b>	<b>BIC</b>	<b>Pred. LL</b>	<b>Notes</b>
<b>LCCM</b>								Identification Issues
<b>GBM-LCCM Full Covariance</b>	2	-18,506.50	-2,769.56	0.19	5,613.12	5,870.78	-1,336.19	
<b>GBM-LCCM Tied Covariance</b>	2	-18,508.80	-2,769.17	0.61	5,610.34	5,861.03	-1,335.37	

a: joint log-likelihood of the GBMLCCM model (Equation 21)

b: marginal log-likelihood of the GBMLCCM (Equation 31) and the LCCM (Equation 12)

c: Marginal log-likelihood (LL) variance of the different runs

### 4.3.2. AUB Case Study

#### 4.3.2.1. Model Specification

In this application, we only consider continuous variables for clustering in order to investigate the impact of the different covariance structures of GMM. We model the weekly frequency of commuting by three different modes (shared-taxi ‘ST’, shuttle ‘SH’, and current mode ‘Car’). The choice variables are then multivariate counts of commuting by three modes during a week with the total count (total number of weekly trips) being fixed to five as the number of times an individual commutes to the university is expected to be rather exogenous than endogenous due to institutional constraints on schedule. Multivariate count data with a fixed total count can be modeled by using a full enumeration of all combinations approach (Sfeir et al., 2020) where the choice set contains all possible combinations of weekly mode frequencies. As such, an alternative is defined as the number of weekly trips an individual would conduct by each of the available modes. In such an approach, the universal choice set would consist of all possible combinations of weekly frequencies of using the three available modes. Knowing that three travel modes are available (ST, SH, and Car) and the sample contains people who commute five days per week to AUB, the choice set consists of 21 alternatives. The systematic utility of an individual  $n$  choosing a specific combination of weekly frequency of three modes  $(ST_n, SH_i, Car_j)^6$  during time period (or scenario)  $t$ , conditional on her/him belonging to class  $k$  can then be specified as follows:

---

<sup>6</sup> For instance,  $(ST_2, SH_1, Car_3)$  means that an individual chose to commute during a specific week, twice by shared taxi (ST), once by shuttle (SH), and three times by car. This combination corresponds to one alternative. There are 21 possible combinations, hence 21 alternatives.

$$\begin{aligned}
V_{n(ST_h,SH_i,Car_j)t|k} &= C_{ST_h,k} + C_{SH_i,k} + C_{Car_j,k} \\
&+ h \times (\beta_{Cost_{ST,k}} Cost_{n,ST,t} + \beta_{TT_{ST,k}} TT_{n,ST,t}) \\
&+ i \times (\beta_{Cost_{SH,k}} Cost_{n,SH,t} + \beta_{TT_{SH,k}} TT_{n,SH,t} + \beta_{Head,k} Head_{n,SH,t}) \\
&+ j \times (\beta_{Cost_{Car,k}} Cost_{n,Car,t} + \beta_{TT_{Car,k}} TT_{n,Car,t}),
\end{aligned} \tag{33}$$

where  $h$ ,  $i$ , and  $j$  are values between 0 and 5 that represent the number of weekday trips by shared-taxi, shuttle and car, respectively. It is assumed that the impact of travel cost, travel time, and headway variables on the utility is proportional to the number of weekly trips by each mode ( $h$ ,  $i$ , and  $j$ ). Moreover, the travel cost and travel time coefficients are specified as mode-specific. The  $C$ 's are constants related to the weekly frequency of the three modes and replace the traditional alternative-specific constants (ASCs) that are defined for each alternative (Ben-Akiva & Abou-Zeid, 2013; Sfeir et al., 2020). Six constants need to be defined for each of the three modes (ST, SH, and Car) since the number of times each mode can be selected per week varies between 0 and 5. Finally, four constants ( $C_{ST_0}$ ,  $C_{SH_0}$ ,  $C_{Car_0}$ ,  $C_{Car_5}$ ) are fixed to zero for identification purposes.

The latent classes of the two models (LCCM and GM-LCCM) are characterized by socio-economic variables while the class-specific utility functions of each alternative are characterized by the corresponding travel time, travel cost and constants related to the frequency of using the available modes. Table 8 shows the explanatory variables used in the two components of the two models, LCCM and GM-LCCM. Several other variables such as income, household car ownership, and parking location were tested but they were insignificant. The coefficients of travel cost and travel time are specified as alternative (mode)-specific to account for variations in Values of Time (VOT) across users of different modes (Guevara, 2017).

Table 8: Explanatory variables used in the models

Variable	Type	Description	Sub-Model
Cost <sub>ST</sub>	Continuous variable	Cost of a one-way trip by shared-ride taxi (in 1,000 L.L.) <sup>7</sup>	Class-specific choice model
Cost <sub>SH</sub>	Continuous variable	Cost of a one-way trip by shuttle including parking cost at the satellite parking (in 1,000 L.L.)	
Cost <sub>Car</sub>	Continuous variable	Fuel and parking cost of a one-way trip by car (in 1,000 L.L.)	
TT <sub>ST</sub>	Continuous variable	Travel time of one-way trip by shared taxi (in hours)	
TT <sub>SH</sub>	Continuous variable	Travel time of one-way trip by shuttle including access time to the satellite parking (in hours)	
TT <sub>Car</sub>	Continuous variable	Travel time of one-way trip by car (in hours)	
Headway	Continuous variable	Shuttle headway (in hours)	
Age	Continuous variable	Age of the respondent (in years/10)	Class membership model
Grade	Continuous variable	A number between 1 and 16 used to specify the job, seniority, and salary of a staff member (Grade/10)	
C/D	Continuous variable	Ratio of number of cars available over number of licensed drivers per household	
Nb	Continuous variable	Number of people who are usually present in the car during the trip from home to AUB	

<sup>7</sup> 1 USD = 1,500 Lebanese Lira (L.L.) at the time the survey was conducted.

#### 4.3.2.2. Estimation Results

Table 9 presents summary statistics of the LCCM and GM-LCCM. For the LCCM, it was not possible to increase the number of latent classes beyond three. In doing so, the LCCM generated very high standard errors for the class-specific parameter estimates. As for the GM-LCCM, there were no identification problems involved in increasing the number of latent classes up to five. However, GM-LCCM with higher number of latent classes ( $K > 2$ ) resulted in positive travel cost and/or travel time coefficients, except for the spherical structure model with three latent classes, and thus these models are excluded from the comparison. Note that the full sub-sample, consisting of 650 respondents and 2,600 choice observations, is used for estimation. The predictive power of the models is compared using the 5-fold cross validation technique. The dataset is divided into 5 subsets and each model is trained 5 times. Each time, the models are trained on 4 different subsets and tested on the remaining one. Next, the log-likelihood of each of the test sets is calculated and the average value is reported. For the case of two latent classes ( $K = 2$ ), results show that the tied structure model has similar marginal log-likelihood as the LCCM but a better prediction log-likelihood. This suggests that the GM-LCCM performs better in terms of prediction accuracy although both models have similar goodness-of-fit measure (LL). The three other covariance structures (full, diagonal, and spherical) have also a better prediction accuracy than the LCCM.

Tables 10 and 11 present estimates of the sub-models of LCCM and tied-GM-LCCM with two classes in addition to the VOT estimates (values between parentheses are p-values). The covariance estimates are not shown for conciseness. Results show that the estimates of the class-specific choice models of the two approaches are almost

the same. All travel cost and travel time parameters have the expected negative sign. Members of the second class seem to be more sensitive to travel time. Next, the latent classes are described.

*K1: old with high car ownership*

The class membership results of the LCCM reveal that members of the first class are more likely old individuals and staff with high grades who live in households with high car ownership. The signs of the means from the class membership model of the GM-LCCM lead to the same conclusion.

*K2: Young with low car ownership*

On the contrary, results of the class membership model of the LCCM reveal that members of the second class are more likely young people and staff with low grades who live in households with fewer cars, and do not share rides to AUB (although the C/D and Nb variables are insignificant). The signs of the means from the class membership model of the GM-LCCM also lead to the same conclusion.

Members of the first class have similar VOT for car and shuttle, which is also a trip by car where a user parks his/her car in a parking garage and uses the shuttle as a first/last mile service to/from AUB, while members of the second class have higher VOT for car. In terms of log-likelihood, both models have the same fitted value. We believe that the improvement in prediction accuracy (Table 9) is due to the changes in the class membership model since the parameter estimates of both class-specific choice models are almost the same (Tables 10 and 11).

Moreover, results of the GM-LCCM with three latent classes and a spherical covariance structure are presented in Table 12. Individuals from the third class appear to

be insensitive towards travel cost of car and travel time of shuttle, hence the high and low VOTs of car and shuttle, respectively. Next, the three latent classes are described.

*K1: The oldest*

The first class includes the oldest individuals (highest  $\mu_{age}$  across all classes) and staff with high grades who live in households with moderate car ownership.

*K2: Young with low car ownership*

The second class is characterized by young individuals ( $\mu_{age} < 0$ ) and staff with low grades ( $\mu_{Grade} < 0$ ) who belong to households with low car ownership.

*K3: The youngest with highest car ownership*

The third class has the youngest individuals (lowest  $\mu_{age}$  across the three classes) who live in households with high car ownership (highest  $\mu_{C/D}$  across the three classes).

Going back to Table 9, it is clear that the GM-LCCM with three latent classes has better joint LL, marginal LL, AIC, and average prediction LL, than both LCCM and GM-LCCM with two latent classes. However, it comes as no surprise that the LCCM with two latent classes has the lowest BIC. This is due to the nature of the GMM and its different covariance structures which result in higher number of parameters for the proposed GM-LCCM. Finally, details of all the estimated models are presented in Appendix B.

Table 9: Summary results of LCCM and GM-LCCM

	Covariance Type	Nb of Parameters	Joint LL <sup>a</sup>	LL <sup>b</sup>	AIC	BIC	Pred. LL
LCCM (K=2)		47		-4,910.92	9,915.84	<b>10,191.41</b>	-1,024.93
GM-LCCM (K=2)	Full	71	-8,476.35	-4,937.64	10,017.28	10,433.57	-1,018.10
	Tied	61	-8,533.22	-4,911.08	9,944.16	10,301.82	-1,012.62
	Diagonal	59	-8,564.64	-4,935.51	9,989.02	10,334.95	-1,017.87
	Spherical	53	-8,575.90	-4,927.54	9,961.08	10,271.83	-1,016.51
GM-LCCM (K=3)	Spherical	80	<b>-7,042.21</b>	<b>-4,893.29</b>	<b>9,946.58</b>	10,415.64	<b>-998.41</b>

a: joint log-likelihood of the GBMLCCM model (Equation 21)

b: marginal log-likelihood of the GBMLCCM (Equation 31) and the LCCM (Equation 12)



Table 10: LCCM – K =2

Table 11: GM-LCCM – K =2

Parameter	Class 1 Class-specific choice model	Class 2 Class-specific choice model	Parameter	Class 1 Class-specific choice model	Class 2 Class-specific choice model
C <sub>car1</sub>	-2.56 (0.00)	0.372 (0.00)	C <sub>car1</sub>	-2.50 (0.00)	0.361 (0.00)
C <sub>car2</sub>	-2.06 (0.00)	0.298 (0.01)	C <sub>car2</sub>	-2.04 (0.00)	0.290 (0.01)
C <sub>car3</sub>	-2.36 (0.00)	0.516 (0.00)	C <sub>car3</sub>	-2.39 (0.00)	0.508 (0.00)
C <sub>car4</sub>	-3.09 (0.00)	-0.422 (0.01)	C <sub>car4</sub>	-3.08 (0.00)	-0.430 (0.00)
C <sub>ST1</sub>	-1.60 (0.03)	-0.464 (0.00)	C <sub>ST1</sub>	-1.62 (0.04)	-0.465 (0.00)
C <sub>ST2</sub>	-2.10 (0.00)	-0.172 (0.24)	C <sub>ST2</sub>	-2.09 (0.00)	-0.174 (0.24)
C <sub>ST3</sub>	-1.05 (0.03)	-0.108 (0.61)	C <sub>ST3</sub>	-1.08 (0.03)	-0.108 (0.61)
C <sub>ST4</sub>	-3.08 (0.00)	-0.347 (0.25)	C <sub>ST4</sub>	-3.15 (0.00)	-0.347 (0.25)
C <sub>ST5</sub>	-0.158 (0.53)	-0.217 (0.53)	C <sub>ST5</sub>	-0.159 (0.53)	-0.209 (0.55)
C <sub>SH1</sub>	-2.256 (0.00)	-0.280 (0.02)	C <sub>SH1</sub>	-2.30 (0.00)	-0.286 (0.02)
C <sub>SH2</sub>	-2.99 (0.00)	0.413 (0.00)	C <sub>SH2</sub>	-3.03 (0.00)	0.403 (0.00)
C <sub>SH3</sub>	-2.26 (0.00)	0.678 (0.00)	C <sub>SH3</sub>	-2.29 (0.00)	0.661 (0.00)
C <sub>SH4</sub>	-3.93 (0.00)	0.373 (0.09)	C <sub>SH4</sub>	-4.02 (0.00)	0.354 (0.11)
C <sub>SH5</sub>	-1.52 (0.00)	0.378 (0.16)	C <sub>SH5</sub>	-1.52 (0.00)	0.379 (0.15)
Cost <sub>Car</sub>	-0.0446 (0.00)	-0.0456 (0.00)	Cost <sub>Car</sub>	-0.0442(0.00)	-0.0462 (0.00)
Cost <sub>ST</sub>	-0.101 (0.00)	-0.109 (0.00)	Cost <sub>ST</sub>	-0.101 (0.00)	-0.110 (0.00)
Cost <sub>SH</sub>	-0.0400 (0.00)	-0.0998 (0.00)	Cost <sub>SH</sub>	-0.0401 (0.00)	-0.0993 (0.00)
TT <sub>Car</sub>	-0.409 (0.00)	-0.658 (0.00)	TT <sub>Car</sub>	-0.409 (0.00)	-0.653 (0.00)
TT <sub>ST</sub>	-0.372 (0.00)	-0.646 (0.00)	TT <sub>ST</sub>	-0.372 (0.00)	-0.641 (0.00)
TT <sub>SH</sub>	-0.252 (0.00)	-0.387 (0.00)	TT <sub>SH</sub>	-0.252 (0.00)	-0.384 (0.00)
Headway	-0.0423 (0.65)	-0.565 (0.00)	Headway	-0.0442 (0.64)	-0.561 (0.00)
Parameter	Class membership model		Parameter	Class membership model	
ASC	-	2.27 (0.00)	$\pi$	0.575	0.425
Age	-	-0.587 (0.00)	$\mu_{Age}$	0.303	-0.409
Grade	-	-0.569 (0.00)	$\mu_{Grade}$	0.225	-0.303
C/D	-	-0.267 (0.37)	$\mu_{C/D}$	0.0459	-0.062
Nb	-	-0.0850 (0.26)	$\mu_{Nb}$	0.0513	-0.0693
Mode	VOT (\$/hr)		Mode	VOT (\$/hr)	
Car	6.11	9.61	Car	6.16	9.42
ST	2.44	3.96	ST	2.45	3.90
SH	4.20	2.59	SH	4.19	2.58

Cost variables are in 1,000 L.L.  
Travel Time and Headway variables are in hours

Cost variables are in 1,000 L.L.  
Travel Time and Headway variables are in hours

Table 12: GM-LCCM - K = 3

Parameter	Class 1	Class 2	Class 3
	Class-specific choice model		
C <sub>car1</sub>	-2.24 (0.00)	0.444 (0.00)	-0.102 (0.78)
C <sub>car2</sub>	-1.82 (0.00)	0.290 (0.02)	0.327 (0.25)
C <sub>car3</sub>	-2.11 (0.00)	0.677 (0.00)	-0.0412 (0.88)
C <sub>car4</sub>	-2.90 (0.00)	-0.224 (0.18)	-1.26 (0.00)
C <sub>ST1</sub>	-1.61 (0.01)	-0.331 (0.01)	-1.11 (0.00)
C <sub>ST2</sub>	-2.31 (0.00)	-0.00930 (0.95)	-0.907 (0.03)
C <sub>ST3</sub>	-1.07 (0.02)	0.0262 (0.91)	-0.938 (0.09)
C <sub>ST4</sub>	-3.23 (0.00)	-0.073 (0.82)	-1.78 (0.08)
C <sub>ST5</sub>	-0.181 (0.47)	-0.284 (0.48)	-0.24 (0.76)
C <sub>SH1</sub>	-2.19 (0.00)	-0.240 (0.07)	-0.557 (0.08)
C <sub>SH2</sub>	-3.17 (0.00)	0.556 (0.00)	-0.535 (0.11)
C <sub>SH3</sub>	-2.12 (0.00)	0.862 (0.00)	-0.959 (0.03)
C <sub>SH4</sub>	-4.38 (0.00)	0.627 (0.01)	-1.63 (0.00)
C <sub>SH5</sub>	-1.45 (0.00)	0.601 (0.04)	-1.15 (0.05)
Cost <sub>Car</sub>	-0.0451 (0.00)	-0.0707 (0.00)	-0.0172 (0.11)
Cost <sub>ST</sub>	-0.0991 (0.00)	-0.107 (0.00)	-0.123 (0.00)
Cost <sub>SH</sub>	-0.0421 (0.00)	-0.0832 (0.00)	-0.120 (0.00)
TT <sub>Car</sub>	-0.410 (0.00)	-0.717 (0.00)	-0.614 (0.00)
TT <sub>ST</sub>	-0.384 (0.00)	-0.777 (0.00)	-0.349 (0.01)
TT <sub>SH</sub>	-0.259 (0.00)	-0.519 (0.00)	-0.107 (0.26)
Headway	-0.0114 (0.91)	-0.757 (0.00)	-0.380 (0.06)
Variable	Class membership model		
$\pi$	0.570	0.339	0.091
$\mu_{Age}$	0.329	-0.314	-0.897
$\mu_{Grade}$	0.258	-0.172	-0.981
$\mu_{C/D}$	0.0456	-0.222	0.540
$\mu_{Nb}$	0.0778	0.0861	-0.810
Mode	VOT (\$/hr)		
Car	6.07	6.76	23.81
ST	2.58	4.85	1.89
SH	4.10	4.16	0.60

Cost variables are in 1,000 L.L.

Travel Time and Headway variables are in hours

#### 4.4. Conclusion

In this chapter, we investigated the feasibility of combining Gaussian-Bernoulli Mixture Models with Latent Class Choice Models. The model was tested and compared to the traditional LCCM using a revealed preferences case study on travel mode choice behavior. The model was also tested and compared to LCCM using a stated preferences case study on weekly frequencies of commuting by different modes. Results showed that the GBM-LCCM is capable of capturing more complex taste heterogeneity than the traditional LCCM by identifying a larger number of latent classes. This might be due to the fact that mixture models allow more flexibility than the linear-in-parameters utility specification of the latent classes. In addition, it is capable of improving the prediction accuracy of the choice models. These improvements are accomplished without any interpretability losses, neither at the class membership level nor at the class-specific choice model level. In fact, the latent classes can be easily interpreted and marginal effects in addition to economic indicators (e.g., willingness to pay) can be directly inferred from the model. To sum up, this new approach satisfies the main properties of an effective econometric behavioral model, as set by McFadden.

However, the proposed model and the applications presented in this chapter are not devoid of limitations. There are several extensions that could be explored further. First, the Gaussian-Bernoulli mixture model assumes that the continuous-binary variables that are used for clustering are uncorrelated. Although the Gaussian part of the mixture model offers different covariance structures for the continuous set of variables, the proposed model should be extended to capture correlations between all continuous-binary variables of the class membership model. This will be addressed by the second proposed model that will be presented in the next Chapter. A second straightforward

extension could be related to within-class heterogeneity. Previous studies have shown that individuals with similar socio-economic characteristics, and thus belonging to the same latent class, might not have the same preferences or taste homogeneity (Bujosa et al., 2010). Therefore, a natural extension of the GBM-LCCM is to integrate random distributions or mixture of random distributions of taste coefficients within the class-specific choice models. Third, although two different types of datasets have been used and several specifications in addition to a logarithmic transformation have been tested, it would be worthwhile to investigate whether the findings of this chapter generalize to different applications, specifications, and attribute transformations.

## CHAPTER 5

### GAUSSIAN PROCESS LATENT CLASS CHOICE MODEL

This chapter develops a Gaussian Process – Latent Class Choice Model (GP-LCCM) by incorporating a Gaussian Process (GP) into the LCCM structure to allow for more complex and flexible discrete representation of heterogeneity and as a result to improve the overall model fit and prediction accuracy compared to the standard LCCM. Moreover, Gaussian Processes allow us to overcome the continuous-binary limitation of the Gaussian-Bernoulli mixture model. The GP-LCCM framework makes use of Gaussian Processes to replace the class membership component of the traditional LCCM. The proposed model would rely on GPs as a nonparametric component to probabilistically divide the population into behaviorally homogenous classes while simultaneously relying on random utility models to develop class-specific choice models. We derive and implement an Expectation-Maximization (EM) algorithm for training a Gaussian Process classification approach as a clustering tool while concurrently learning the parameter estimates of the class-specific choice models. By doing so, we contribute to the discrete choice modeling literature by formulating, to the author’s knowledge, the first Gaussian Process choice model within an LCMM framework, thereby allowing for more modeling flexibility and higher prediction accuracy. We also develop a Gaussian Process model for clustering by incorporating the Laplace approximation approach (Williams & Barber, 1998), which is used for Gaussian process classification problems, in an iterative EM algorithm.

We start by presenting the Gaussian Process formulation for classification problems (e.g., prediction of class labels) (Section 5.1). Next, we present different

covariance kernel functions for a Gaussian Process (Section 5.2). Finally, we combine the concepts of LCCM (Section 3.1), GP (Section 5.1) and kernels (Section 5.2) to define the Gaussian Process – Latent Class Choice Model and derive an Expectation-Maximization (EM) algorithm for estimation (Section 5.3).

## 5.1. Gaussian Process

Gaussian Processes are a powerful and flexible probabilistic machine learning technique (Rasmussen & Williams, 2006) that instead of parameterizing the target variables (e.g., class labels) or placing priors over the unknown parameters of a predefined distribution (e.g., mean and variance of a normal distribution), define priors over latent functions directly (Mackay, 2003; Rasmussen & Williams, 2006). It can be considered as a generalization of a Gaussian distribution over a finite vector space to an infinite function space (Mackay, 2003). Therefore, a GP is specified by a mean function and a covariance function usually known as kernel.

For the sake of simplicity and without loss of generality, we consider a binary case ( $K = 2$ ) where the training data consists of  $S$ , a matrix of  $N$  vectors  $S_n$  (vector of characteristics of decision-maker  $n$  with a dimension equal to  $D_S$ ), and  $q_k$  a vector of  $N$  target outputs  $q_{nk}$  (class label, equal to 1 or 0). The goal is to model the posterior distribution of the target outputs by defining a prior distribution over a latent function  $f$  by using a multivariate Gaussian distribution with a mean function  $m(S_n)$ , that represents the expected value for each latent variable  $f(S_n)$ , and a kernel (covariance function)  $C(S_n, S_m) = cov[f(S_n), f(S_m)]$ , that represents the variance between every pair of latent variables  $f(S_n)$  and  $f(S_m)$ . Note that it is common to specify a GP with a zero mean function without loss of generality (Rasmussen & Williams, 2006). A GP

prior is therefore specified for the function values  $f$ ,  $f \sim GP(m(S_n) = 0, C(S_n, S_m))$  such that:

$$p(f|S) = \mathcal{N}(0, C), \quad (34)$$

where  $f$  is a vector of  $N$  latent variable values  $f_n$ ,  $S$  is a matrix of  $N$  vectors of  $S_n$  and  $C$  is a  $(N \times N)$  covariance matrix defined by a covariance function (or kernel) such that  $C_{n,m} = C(S_n, S_m)$ . Note that  $C$  could be a group of  $D_S$   $(N \times N)$  matrices in case an Automatic Relevance Determination (ARD) covariance function is used (Refer to section 5.2).

The next step is to specify an appropriate likelihood or link function for the classes to obtain a probabilistic classification since the target outputs are discrete (0 or 1). The link function could be a sigmoid function or a cumulative density function of a standard normal distribution. We make use of a sigmoid function as follows:

$$P(q_{nk}|f_n) = \frac{1}{1 + \exp(-f_n)}. \quad (35)$$

Then, the posterior over  $f$  can be determined using Bayes' theorem as follows:

$$P(f_n|q_{nk}, S_n) = \frac{P(q_{nk}|f_n)P(f_n|S_n)}{P(q_{nk}|S_n)}. \quad (36)$$

The combination of a Gaussian Process prior with a non-Gaussian link function results in a non-Gaussian posterior that is analytically intractable. Nevertheless, the posterior can be approximated by a GP using different approximation techniques such as Markov Chain Monte Carlo (MCMC) (Neal, 1999), Variational Inference (VI) (Gibbs & Mackay, 2000) and Expectation Propagation (EP) (Minka, 2001; Opper & Winther, 2000). In this research, we make use of the Laplace approximation (C. Bishop, 2006; Rasmussen & Williams, 2006; Williams & Barber, 1998) which approximates the posterior with a Gaussian by taking a second-order Taylor expansion of the logarithm of

the posterior around its maximum. For more details about the Laplace approximation, readers may refer to Rasmussen and Williams (2006, sec. 3.4) and Bishop (2006, sec. 6.4.6).

## 5.2. Kernels

The choice of a suitable covariance function (kernel) is a crucial step in learning a Gaussian Process that generalizes beyond the training data since the kernel can shape the distribution we wish to learn in different ways and determine the characteristics of the fitted function (such as smoothness, periodicity, stationarity and isotropy). Different kernels or combinations (addition or multiplication) of kernels can be used to generate more complex structures and improve the GP flexibility. We present next the most common kernels and the ones that are used in our applications (Chapter 6).

### 5.2.1. Squared Exponential Kernel (SE) / Radial Basis Function (RBF)

The most common choice of kernel is the Squared Exponential kernel (SE), also known as Radial Basis Function (RBF), which is defined as follows:

$$k_{SE}(S_n, S_m) = \lambda^2 \exp\left(-\frac{r^2}{2\ell^2}\right), \quad (37)$$

where  $r = |S_n - S_m|$  is the Euclidean distance between two observations  $S_n$  and  $S_m$  (e.g., characteristics of two individuals  $n$  and  $m$ ),  $\lambda^2$  is the variance of the distance between two observations and  $\ell$  is the length-scale which determines the smoothness of the kernel function and the importance of the features (independent variables).

The SE kernel is a stationary kernel that is infinitely differentiable (mean square derivatives of all orders) and as such is very smooth. However, it is believed that such



strong smoothness is unrealistic for some applications and the Matérn kernel is instead recommended (Stein, 1999).

### 5.2.2. Matérn Kernel

The Matérn kernel is a stationary kernel that can be considered as a generalization of the SE kernel. It is defined as follows:

$$k_{Matern}(S_n, S_m) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}r}{\ell} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu}r}{\ell} \right), \quad (38)$$

where  $r = |S_n - S_m|$  is the Euclidean distance between two observations  $S_n$  and  $S_m$  (e.g., characteristics of two individuals  $n$  and  $m$ ),  $\nu$  is a positive parameter that controls the smoothness of the function (lower values result in less smooth functions),  $\ell$  is the length-scale of the kernel,  $\Gamma$  is the gamma function, and  $K_\nu$  is a modified Bessel function (Abramowitz & Stegun, 1965; Mackay, 1998).

The most interesting and commonly used cases for machine learning are  $\nu = 3/2$  and  $\nu = 5/2$ .

It is to be noted that both of the above kernels could be used with Automatic Relevance Determination (ARD) by specifying the length-scale as a vector of dimension  $D_S$  equal to the dimension of  $S_n$  (Rasmussen and Williams, 2006). Large length-scale values mean the function values are uncorrelated and the corresponding feature(s) should be removed from the model (Rasmussen and Williams, 2006).

Other kernel functions can be used such as periodic, exponential, radial quadratic, or piecewise polynomial, to name a few. For more details, readers may refer to Rasmussen and Williams (2006, sec. 4).

### 5.3. Gaussian Process – Latent Class Choice Model

We now present the formulation of the Gaussian Process – Latent Class Choice Model (GP-LCCM). Similar to the traditional LCCM, the GP-LCCM consists of two components, a class membership model and a class-specific choice model. The former is defined as a Gaussian Process that probabilistically assigns decision-makers to behaviorally homogeneous latent classes/clusters, while the latter formulates class-specific choice probabilities using typical discrete choice models (e.g., MNL). Figure 5 shows the graphical representation of the proposed model. Hatched circles represent observed variables and choices while white circles symbolize unknown parameters and latent variables.

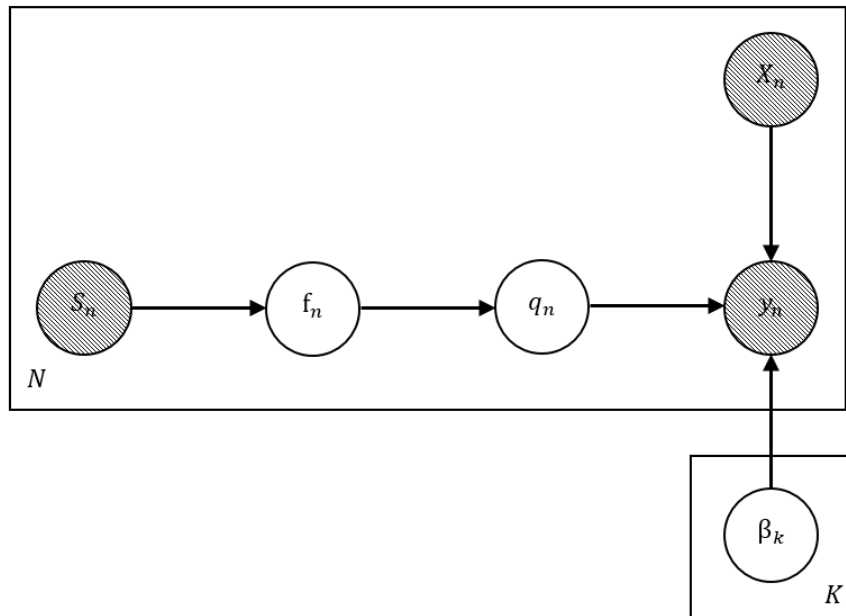


Figure 5: Graphical representation of the proposed Gaussian Process – Latent Class Choice Model (GP-LCCM) for a set of  $N$  decision-makers and  $K$  clusters/latent classes

#### 5.3.1. Proposed Model

Given the conditional independence properties of the graphical model structure of the GP-LCCM (Figure 5), the joint probability of  $f_n$ ,  $y_n$ , and  $q_{nk}$  can be formulated

as the product of the Gaussian prior (Equation 34 and first term on the right hand side below), the link function or the likelihood of  $q_{nk}$  conditional on the latent function  $f_n$  (Equation 35 and second term on the right hand side below) and the choice probability conditional on the class (Equation 10 and third term on the right hand side below), as follows:

$$P(f_n, y_n, q_{nk} = 1 | S_n, X_n, \beta_k) = P(f_n | S_n) P(q_{nk} = 1 | f_n) P(y_n | X_n, q_{nk} = 1, \beta_k). \quad (39)$$

The joint probability of  $f_n$  and  $y_n$  is then obtained by summing Equation (39) over all classes  $K$  (we omit the dependencies on the left hand side of the equation to make the notation less cluttered):

$$\begin{aligned} P(f_n, y_n) &= \sum_{k=1}^K P(f_n, y_n, q_{nk} = 1 | S_n, X_n, \beta_k) \\ &= \sum_{k=1}^K P(f_n | S_n) P(q_{nk} = 1 | f_n) P(y_n | X_n, q_{nk} = 1, \beta_k). \end{aligned} \quad (40)$$

Finally, the joint likelihood function of the GP-LCCM model for a sample of  $N$  decision-makers is given by:

$$\begin{aligned} P(f, y) &= \prod_{n=1}^N P(f_n, y_n) = \prod_{n=1}^N \sum_{k=1}^K P(f_n | S_n) P(q_{nk} = 1 | f_n) \\ &\quad \times \prod_{t=1}^{T_n} \prod_{j=1}^J P(y_{njt} = 1 | X_{njt}, q_{nk} = 1, \beta_k)^{y_{njt}}, \end{aligned} \quad (41)$$

To learn the parameters  $\beta_k$  and the hyper-parameters of the kernel, the log of the above likelihood should be maximized and evaluated over the unknown parameters. Similarly to the GBM-LCCM, an EM-based algorithm is derived and implemented for estimating the proposed GP-LCCM.

### 5.3.2. EM Algorithm

The EM algorithm requires writing the likelihood function (Equation 41) assuming that the class assignments ( $q_{nk}$ ) are no longer latent:

$$P(f, y) = \prod_{n=1}^N \prod_{k=1}^K [P(f_n | S_n) P(q_{nk} = 1 | f_n)]^{q_{nk}} \times \prod_{n=1}^N \prod_{k=1}^K \prod_{t=1}^{T_n} \prod_{j=1}^J P(y_{njt} = 1 | X_{njt}, q_{nk} = 1, \beta_k)^{y_{njt} q_{nk}}. \quad (42)$$

The logarithm of the above likelihood is then the sum of the two sub-components of the model, the class membership component (first term on the right-hand side below) and the class-specific choice component (second term), as follows:

$$LL = \sum_{n=1}^N \sum_{k=1}^K q_{nk} \log [P(f_n | S_n) P(q_{nk} = 1 | f_n)] + \sum_{n=1}^N \sum_{k=1}^K \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} q_{nk} \log [P(y_{njt} = 1 | X_{njt}, q_{nk} = 1, \beta_k)]. \quad (43)$$

Next, the unknown choice model parameters ( $\beta_k$ ) can be estimated by setting the derivatives of the log-likelihood (Equation 43) with respect to the unknown parameters to zero if and only if  $q_{nk}$ 's are known. Similarly, the hyper-parameters of the GP kernel function can be found using the Laplace approximation method if and only if  $q_{nk}$ 's are known. Therefore, the expectation of  $q_{nk}$  (E-step) is calculated using Bayes' theorem as follows:

$$P(q_{nk} = 1 | y_n, S_n, f_n, X_n, \beta_k) \propto P(q_{nk} = 1 | f_n, S_n) P(y_n | X_n, q_{nk} = 1, \beta_k), \quad (44)$$

$$E[q_{nk}] = \gamma_{q_{nk}} = \frac{P(q_{nk} = 1 | f_n, S_n) P(y_n | X_n, q_{nk} = 1, \beta_k)}{\sum_{c=1}^K P(q_{nc} = 1 | f_n, S_n) P(y_n | X_n, q_{nc} = 1, \beta_c)}. \quad (45)$$

Then, the expected value of the log-likelihood w.r.t  $q_{nk}$  is maximized instead of Equation 43, due to the unknown values of the latent variables  $q_{nk}$ , to find/update the unknown hyper/parameters (M-step). The expected log-likelihood function is given by:

$$\begin{aligned}
E[LL] = & \sum_{n=1}^N \sum_{k=1}^K \gamma_{q_{nk}} \log[P(f_n|S_n)P(q_{nk} = 1|f_n)] \\
& + \sum_{n=1}^N \sum_{k=1}^K \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} \gamma_{q_{nk}} \log[P(y_{njt} = 1|X_{njt}, q_{nk} = 1, \beta_k)].
\end{aligned} \tag{46}$$

Setting the derivative of the above expected log-likelihood with respect to  $\beta_k$  to zero, we can find the updated solution of  $\beta_k$  as follows:

$$\beta_k = \underset{\beta_k}{\operatorname{argmax}} \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{j=1}^J y_{njt} \gamma_{q_{nk}} \log \left[ \frac{e^{X'_{njt} \beta_k}}{\sum_{j'=1}^J e^{X'_{njt} \beta_k}} \right]. \tag{47}$$

Note that closed-form solutions cannot be obtained for Equation 47. Instead, we rely on the gradient-based numerical optimization method BFGS (Nocedal & Wright, 2006) or the constrained version L-BFGS-B (C. Zhu et al., 1997).

As for estimating the hyper-parameters of the Gaussian Process kernel, a Laplace approximation method is applied. However, the target variables (class labels) should be known since the Laplace approximation is applied to classification applications. For the sake of simplicity and without loss of generality, we consider the case of two classes ( $K = 2$ ). After calculating the expectations of  $q_{nk}$ 's in the E-step (which are continuous values between 0 and 1), class labels are generated using hard clustering/assignment as follows: if  $E[q_{n0}] > E[q_{n1}]$ , then individual  $n$  belongs to class 0, otherwise individual  $n$  belongs to class 1. Next, a Laplace approximation is applied and the hyper-parameters of the kernel are estimated.

The steps of the EM estimation of the proposed GP-LCCM with two classes

( $K = 2$ ) are:

1. Initialize the parameters  $\beta_k$  and assign each individual to a class (0 or 1) randomly
2. Select a kernel function and initialize the corresponding hyper-parameters
3. **E-step:** Estimate the expectations of  $q_{nk}$ 's using Equation 45
4. **M-step:**
  - i. Re-estimate/update the parameters  $\beta_k$  using Equation 47 and the expectations from the E-step
  - ii. Assign each individual to one class (0 or 1) using the expectations from the E-step as follows: if  $E[q_{n0}] > E[q_{n1}]$ , then individual  $n$  belongs to class 0, otherwise individual  $n$  belongs to class 1
  - iii. Re-estimate/update the hyper-parameters of the kernel function using the Laplace approximation method and the class labels from the previous step (ii)
5. Evaluate the log-likelihood using the current values of the hyper/parameters and check if the convergence criterion is satisfied. If, not return to step 3
6. Finally, after convergence is reached, estimate the marginal probability of observing a vector of choices  $y$ ,  $P(y)$ , for all decision-makers  $N$  as follows:

$$P(y) = \prod_{n=1}^N \sum_{k=1}^K P(q_{nk} = 1 | f_n, S_n) P(y_n | X_n, q_{nk} = 1, \beta_k). \quad (48)$$

The marginal probability (Equation 48) is calculated in order to compare the GP-LCCM with the traditional LCCM (Equation 12) and assess out-of-sample prediction accuracies. For multi-class problems ( $K = 2$ ),  $K$  binary one-versus-rest classifiers are estimated to classify each class against all the rest using the Laplace approximation method. The implementation of the multi-class case is based, within each EM iteration, on Algorithms 3.1, 3.2, and 5.1 from Gaussian Processes for Machine Learning (GPML) by Rasmussen and Williams (2006).

## CHAPTER 6

### GP-LCCM APPLICATIONS

We present three different mode choice applications of the proposed GP-LCCM approach. We benchmark the proposed model against the traditional LCCM and the proposed GBM-LCCM/ GM-LCCM by using the same specifications for all models. Section 6.1 presents the datasets. Section 6.2 describes the implementation of the different models. Section 6.3 presents the model formulations/specifications and estimation results. Section 6.4 concludes.

#### **6.1. Data**

Three different datasets are used to evaluate the GP-LCCM approach. We make use of the AUB dataset (Section 4.1.2), London dataset (Section 4.1.1) and Swissmetro dataset which is presented in Section 6.1.1.

##### ***6.1.1 Swissmetro Dataset***

The third application is based on the famous Swissmetro dataset which consists of SP survey data collected in Switzerland during March of 1998 to assess the potential demand for the formerly proposed Swissmetro, a maglev underground transport system (Bierlaire et al., 2001). Each respondent was offered nine hypothetical scenarios with the following alternatives: Train, Swissmetro (SM) and Car (only for car owners). Each alternative was described by its corresponding attributes such as travel time and travel cost/fare, etc. Socio-economic/demographic information was also collected. For more information, readers may refer to Bierlaire (2018) and Bierlaire et al. (2001). The



original dataset contains 10,728 observations corresponding to 1,192 respondents (751 car users and 441 rail-based travelers). However, observations with missing age, unknown choices and “other” trip purposes are removed. As a result, the used sample consists of 10,692 observations corresponding to 1,188 respondents. The sample is randomly divided into 80% (950 respondents and 8,550 observations) for training/estimation and 20% (238 respondents and 2,142 observations) for testing/prediction.

## 6.2. Implementation

The GP-LCCM is implemented in Python by using some blocks from: 1) the Gaussian Process Classifier (GPC) of the Scikit-Learn library (Pedregosa et al., 2011), which is based on Laplace approximation by Rasmussen and Williams (2006); 2) and `lccm` (El Zarwi, 2017a, 2017b), a python package that implements an EM algorithm for estimating traditional latent class choice models. The code is publicly available on GitHub<sup>8</sup>. The GBM-LCCM/GM-LCCM and traditional LCCM are implemented as mentioned in Section 4.2. Convergence of the three different models is assumed to be reached once the change in the log-likelihood function between two successive EM iterations is smaller than  $1 \times 10^{-4}$ . The GP-LCCMs are estimated five times with different random initialization to assess the stability of the models. All runs are performed on a machine with a core i7 CPU @ 2.40 GHz, 8GB of RAM and a GeForce GT 730M. Note that the socio-economic variables entering the class membership component of the GP-LCCMs are standardized (mean = 0 and standard deviation = 1) prior to the estimation.

---

<sup>8</sup> <https://github.com/gsfair/GP-LCCM>

### **6.3. Applications**

We present in this section the model specifications and estimation results of the AUB, London, and Swissmetro applications in Sections 6.3.1, 6.3.2, and 6.3.3, respectively.

#### ***6.3.1. AUB Case Study***

##### 6.3.1.1. Model Specification

The LCCM and GM-LCCM are specified as in Section 4.3.2. A tied covariance is selected for GM-LCCM with two latent classes and a spherical covariance for GM-LCCM with three latent classes since those specifications proved to generate the best results (Chapter 4 – Table 9). The latent classes and class-specific utility functions of the GP-LCCM are also characterized similarly to LCCM and GM-LCCM (Chapter 4 – Table 8). As for the choice of the kernel function of GPs, and knowing that kernel functions affect the generalization performance of the model, different kernels and combinations of kernels were tested and models with better generalization performance, better in-sample goodness-of-fit measures (LL, AIC, and BIC), and reasonable parameter estimate signs and magnitudes were selected. For the GP-LCCM with two classes, a Matérn kernel with a smoothness parameter ( $\nu$ ) of 2.5 is used while for three latent classes, a combination of a constant and a Matérn kernel with a smoothness parameter of 2.5 is selected. A constant had to be added since a single Matérn kernel resulted in high standard errors for some class-specific parameter estimates.

##### 6.3.1.2. Estimation Results

Summary statistics for the LCCM, GM-LCCM and GP-LCCM are shown in Table 13. The LCCM was only able to identify two latent classes. Increasing the

number of classes beyond two resulted in class-specific estimates with very high standard errors (identification issues). On the other hand, the GM-LCCM and GP-LCCM were able to identify higher number of classes (up to 5). However, such models with more than three latent classes generated positive travel cost and/or travel time coefficients and as such are excluded from the comparison (Table 13). Results show that the GP-LCCM with two classes has better goodness-of-fit measures (Joint LL, LL, AIC, and BIC) and better prediction accuracy in terms of log-likelihood (Pred. LL) than the two other models with two classes. Moreover, the GP-LCCM with three latent classes has better goodness-of-fit and prediction measures than all other models. It is to be noted that, compared to the LCCM with two latent classes, the GP-LCCM with three latent classes improves the in-sample goodness-of-fit (LL) and the out-of-sample prediction accuracy (Pred. LL) by 4.5% and 8.8%, respectively. As for the GM-LCCM with three latent classes, the improvement over the LCCM is less significant with 0.4% for the LL and 2.6% for the Pred. LL (Table 13). It is believed that the superiority of the GP-LCCM stems from the nonparametric nature of GPs, which allows for more flexibility than the parametric structure of the Gaussian mixture models and the linear-in-parameters utility specification of the class membership model of the traditional LCCM. Note that the 5-fold cross validation technique is used to assess the predictive power of the models. Furthermore, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are derived as follows:

$$AIC = 2M - 2 \log LL , \quad (49)$$

$$BIC = M \log D - 2 \log LL , \quad (50)$$

where  $M$  is the number of parameters,  $LL$  is the estimated marginal choice log-likelihood, and  $D$  is the number of data points (observations). For the LCCM and GM-

LCCM, the number of parameters  $M$  is equal to the number of unknown parameters that are statistically estimated using the available data. As for the GP-LCCM, the complexity of the model grows with the amount of data used for estimation since Gaussian Processes are nonparametric models. The true number of parameters would be equal to the number of data points (2600 observations for this application) in addition to the number of choice parameters ( $\beta_k$ ) from the class-specific choice models (42 choice parameters for the case of 2 classes). However, such number would be unreasonable for the estimation of AIC and BIC. Instead, it is common to assume that the number of parameters in a GP model is equal to the number of kernel hyper-parameters (Lloyd et al., 2014; Richter & Toledano-Ayala, 2015). Therefore, for the sake of comparison with other models, we assume that the number of parameters  $M$  for a GP-LCMM is equal to the number of kernel hyper-parameters ( $\nu$  and  $\ell$  in case of a Matérn kernel) in addition to the number of class-specific choice parameters ( $\beta_k$ ).

Table 13: Summary results of the AUB application

<b>K</b>	<b>Model</b>	<b>Nb of parameters</b>	<b>Joint LL<sup>a</sup></b>	<b>LL<sup>b</sup></b>	<b>AIC</b>	<b>BIC</b>	<b>Average Pred. LL</b>
	LCCM	47		-4,910.92	9,915.84	10,191.41	-1,024.93
2	GM-LCCM	61	-8,533.22	-4,911.08	9,944.16	10,301.82	-1,012.62
	GP-LCCM	44	<b>-4,905.31</b>	<b>-4,877.73</b>	<b>9,843.46</b>	<b>10,101.44</b>	<b>-995.76</b>
3	GM-LCCM	80	-7,042.21	-4,893.29	9,946.58	10,415.64	-998.41
	GP-LCCM	72	<b>-4,480.70</b>	<b>-4,691.25</b>	<b>9,526.50</b>	<b>9,948.66</b>	<b>-935.23</b>

a: joint log-likelihood of the GM-LCCM (Equation 21) and GP-LCCM (Equation 44)

b: marginal choice log-likelihood of the LCCM, GM-LCCM (Equation 31) and GP-LCCM (Equation 49)

Estimates of the class-specific choice models of the three LCCMs with two latent classes in addition to the corresponding Values of Time (VOTs) are presented in Table 14 (values between parentheses are p-values). Travel time and travel cost parameter estimates have the expected negative sign. In addition, all parameter

estimates are similar in magnitude and sign with the ones related to LCCM and GM-LCCM being almost the same. The second class is characterized by higher VOT for car while individuals from both classes have, to some extent, similar VOTs for shared-taxi and shuttle. Table 15 shows the estimates of the class-specific choice models and VOTs of the GM-LCCM and GP-LCCM with three latent classes. All travel time and travel cost coefficients have the same order of magnitude and expected negative sign. Moreover, according to the GM-LCCM, individuals belonging to the third class seem to be insensitive towards travel cost of car ( $p$ -value = 0.11) and travel time of shuttle ( $p$ -value = 0.26) while the same variables are highly significant according to the GP-LCCM ( $p$ -values = 0). This insignificance of certain level of service variables in the third class of the GM-LCCM results in very high and low VOTs of car and shuttle, respectively. However, compared to previous mode choice studies of AUB students (Al-Ayyash et al., 2016; Sfeir et al., 2020), the GP-LCCM generates more reasonable estimates of Values of Time than the GM-LCCM. This previous discussion shows that the behavioral and economic interpretability of the class-specific choice models were not jeopardized by the introduction of Gaussian Processes to the LCCM framework. Furthermore, the GP-LCCM is capable of improving the prediction accuracy, capturing more complex heterogeneity than the LCCM since a higher number of classes is identified, and generating more seemingly reliable VOT estimates than the GM-LCCM.

Table 14: Class-specific choice models and VOT (K = 2)

Parameter	LCCM		GM-LCCM		GP-LCCM	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
C <sub>car1</sub>	-2.56 (0.00)	0.372 (0.00)	-2.50 (0.00)	0.361 (0.00)	-2.69 (0.00)	0.351 (0.00)
C <sub>car2</sub>	-2.06 (0.00)	0.298 (0.01)	-2.04 (0.00)	0.290 (0.01)	-1.92 (0.00)	0.282 (0.01)
C <sub>car3</sub>	-2.36 (0.00)	0.516 (0.00)	-2.39 (0.00)	0.508 (0.00)	-2.64 (0.00)	0.483 (0.00)
C <sub>car4</sub>	-3.09 (0.00)	-0.422 (0.01)	-3.08 (0.00)	-0.430 (0.00)	-2.77 (0.00)	-0.449 (0.00)
C <sub>ST1</sub>	-1.60 (0.03)	-0.464 (0.00)	-1.62 (0.04)	-0.465 (0.00)	-1.96 (0.00)	-0.475 (0.00)
C <sub>ST2</sub>	-2.10 (0.00)	-0.172 (0.24)	-2.09 (0.00)	-0.174 (0.24)	-1.83 (0.00)	-0.211 (0.15)
C <sub>ST3</sub>	-1.05 (0.03)	-0.108 (0.61)	-1.08 (0.03)	-0.108 (0.61)	-1.46 (0.02)	-0.0979 (0.63)
C <sub>ST4</sub>	-3.08 (0.00)	-0.347 (0.25)	-3.15 (0.00)	-0.347 (0.25)	-2.24 (0.03)	-0.434 (0.15)
C <sub>ST5</sub>	-0.158 (0.53)	-0.217 (0.53)	-0.159 (0.53)	-0.209 (0.55)	-0.0556 (0.83)	-0.169 (0.62)
C <sub>SH1</sub>	-2.26 (0.00)	-0.280 (0.02)	-2.30 (0.00)	-0.286 (0.02)	-2.83 (0.00)	-0.326 (0.01)
C <sub>SH2</sub>	-2.99 (0.00)	0.413 (0.00)	-3.03 (0.00)	0.403 (0.00)	-2.86 (0.00)	0.350 (0.01)
C <sub>SH3</sub>	-2.26 (0.00)	0.678 (0.00)	-2.29 (0.00)	0.661 (0.00)	-2.47 (0.00)	0.595 (0.00)
C <sub>SH4</sub>	-3.93 (0.00)	0.373 (0.09)	-4.02 (0.00)	0.354 (0.11)	-3.95 (0.00)	0.282 (0.19)
C <sub>SH5</sub>	-1.52 (0.00)	0.378 (0.16)	-1.52 (0.00)	0.379 (0.15)	-1.54 (0.00)	0.367 (0.15)
Cost <sub>Car</sub>	-0.0446 (0.00)	-0.0456 (0.00)	-0.0442(0.00)	-0.0462 (0.00)	-0.0425 (0.00)	-0.0474 (0.00)
Cost <sub>ST</sub>	-0.101 (0.00)	-0.109 (0.00)	-0.101 (0.00)	-0.110 (0.00)	-0.105 (0.00)	-0.108 (0.00)
Cost <sub>SH</sub>	-0.0400 (0.00)	-0.0998 (0.00)	-0.0401 (0.00)	-0.0993 (0.00)	-0.0399 (0.00)	-0.0971 (0.00)
TT <sub>Car</sub>	-0.409 (0.00)	-0.658 (0.00)	-0.409 (0.00)	-0.653 (0.00)	-0.420 (0.00)	-0.630 (0.00)
TT <sub>ST</sub>	-0.372 (0.00)	-0.646 (0.00)	-0.372 (0.00)	-0.641 (0.00)	-0.380 (0.00)	-0.629 (0.00)
TT <sub>SH</sub>	-0.252 (0.00)	-0.387 (0.00)	-0.252 (0.00)	-0.384 (0.00)	-0.255 (0.00)	-0.375 (0.00)
Headway	-0.0423 (0.65)	-0.565 (0.00)	-0.0442 (0.64)	-0.561 (0.00)	-0.0556 (0.56)	-0.562 (0.00)
Mode	VOT (\$/hr)					
Car	6.11	9.61	6.16	9.42	6.59	8.87
ST	2.44	3.96	2.45	3.90	2.42	3.87
SH	4.20	2.59	4.19	2.58	4.26	2.58

Cost variables are in 1,000 L.L.

Travel Time and Headway variables are in hours

Table 15: Class-specific choice models (K = 3)

Parameter	GM-LCCM			GP-LCCM		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
C <sub>car1</sub>	-2.24 (0.00)	0.444 (0.00)	-0.102 (0.78)	-3.54 (0.29)	0.397 (0.00)	0.746 (0.33)
C <sub>car2</sub>	-1.82 (0.00)	0.290 (0.02)	0.327 (0.25)	-2.83 (0.57)	-0.0479 (0.77)	1.56 (0.01)
C <sub>car3</sub>	-2.11 (0.00)	0.677 (0.00)	-0.0412 (0.88)	-2.52 (0.62)	0.140 (0.48)	1.90 (0.00)
C <sub>car4</sub>	-2.90 (0.00)	-0.224 (0.18)	-1.26 (0.00)	-5.20 (0.20)	-1.34 (0.00)	0.347 (0.16)
C <sub>ST1</sub>	-1.61 (0.01)	-0.331 (0.01)	-1.11 (0.00)	-1.33 (0.73)	0.285 (0.07)	-1.45 (0.00)
C <sub>ST2</sub>	-2.31 (0.00)	-0.00930 (0.95)	-0.907 (0.03)	-2.85 (0.57)	0.915 (0.00)	-1.90 (0.00)
C <sub>ST3</sub>	-1.07 (0.02)	0.0262 (0.91)	-0.938 (0.09)	-1.40 (0.78)	1.09 (0.00)	-1.41 (0.03)
C <sub>ST4</sub>	-3.23 (0.00)	-0.073 (0.82)	-1.78 (0.08)	-1.71 (0.61)	0.863 (0.05)	-1.06 (0.23)
C <sub>ST5</sub>	-0.181 (0.47)	-0.284 (0.48)	-0.244 (0.76)	-0.141 (0.60)	2.69 (0.00)	-1.20 (0.06)
C <sub>SH1</sub>	-2.19 (0.00)	-0.240 (0.07)	-0.557 (0.08)	-3.59 (0.267)	0.987 (0.00)	-0.838 (0.00)
C <sub>SH2</sub>	-3.17 (0.00)	0.556 (0.00)	-0.535 (0.11)	-3.48 (0.49)	2.08 (0.00)	-0.812 (0.08)
C <sub>SH3</sub>	-2.12 (0.00)	0.862 (0.00)	-0.959 (0.03)	-1.84 (0.71)	2.18 (0.00)	0.285 (0.68)
C <sub>SH4</sub>	-4.38 (0.00)	0.627 (0.01)	-1.63 (0.00)	-4.36 (0.27)	2.45 (0.00)	-0.231 (0.80)
C <sub>SH5</sub>	-1.45 (0.00)	0.601 (0.04)	-1.15 (0.05)	-1.64 (0.00)	3.17 (0.00)	0.555 (0.35)
Cost <sub>Car</sub>	-0.0451 (0.00)	-0.0707 (0.00)	-0.0172 (0.11)	-0.0451 (0.00)	-0.0645 (0.00)	-0.0524 (0.00)
Cost <sub>ST</sub>	-0.0991 (0.00)	-0.107 (0.00)	-0.123 (0.00)	-0.102 (0.00)	-0.132 (0.00)	-0.101 (0.00)
Cost <sub>SH</sub>	-0.0421 (0.00)	-0.0832 (0.00)	-0.120 (0.00)	-0.0411 (0.00)	-0.118 (0.00)	-0.150 (0.00)
TT <sub>Car</sub>	-0.410 (0.00)	-0.717 (0.00)	-0.614 (0.00)	-0.404 (0.00)	-0.466 (0.00)	-0.680 (0.00)
TT <sub>ST</sub>	-0.384 (0.00)	-0.777 (0.00)	-0.349 (0.01)	-0.360 (0.00)	-0.619 (0.00)	-0.856 (0.00)
TT <sub>SH</sub>	-0.259 (0.00)	-0.519 (0.00)	-0.107 (0.26)	-0.226 (0.00)	-0.319 (0.00)	-0.752 (0.00)
Headway	-0.0114 (0.91)	-0.757 (0.00)	-0.380 (0.06)	-0.0551 (0.59)	-0.381 (0.00)	-0.783 (0.00)
Mode	VOT (\$/hr)					
Car	6.07	6.76	23.81	5.96	4.82	8.65
ST	2.58	4.85	1.89	2.34	3.13	5.64
SH	4.10	4.16	0.60	3.67	1.80	3.35

Cost variables are in 1,000 L.L.

Travel Time and Headway variables are in hours

Results of the class membership model of the LCCM and GM-LCCM with two latent classes are presented in Table 16. According to both the LCCM parameter estimates and the GM-LCCM mean estimates, members of the first class are more likely to be older people and staff with high grades who belong to households with high car ownership and have tendency to share rides to AUB.

Table 16: Class membership estimates of LCCM and GM-LCCM ( $K = 2$ )

Parameter	LCCM			GM-LCCM	
	Class 1	Class 2		Class 1	Class 2
ASC	-	2.27 (0.00)	$\pi$	0.575	0.425
Age	-	-0.587 (0.00)	$\mu_{\text{Age}}$	0.303	-0.409
Grade	-	-0.569 (0.00)	$\mu_{\text{Grade}}$	0.225	-0.303
C/D	-	-0.267 (0.37)	$\mu_{\text{C/D}}$	0.0459	-0.0620
Nb	-	-0.0850 (0.26)	$\mu_{\text{Nb}}$	0.0513	-0.0693

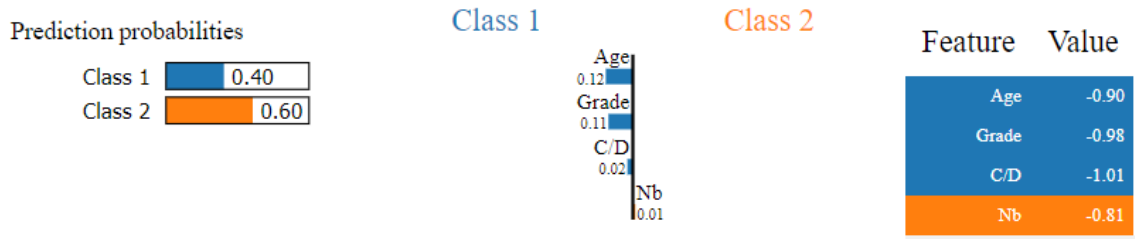
As for the GP-LCCM, the introduction of the nonparametric Gaussian Processes makes the model less transparent at the class membership level. However, the latent classes can still be interpreted although in a different manner than in traditional DCMs and parametric models where interpretability is based on the parameter estimates. Recently, interpretability of machine learning models has become a fundamental area of research and many studies have shown that different techniques can be used for model interpretation (Doshi-velez and Kim, 2017; Ribeiro et al., 2016b, 2016a; Wang et al., 2020, to name a few). One approach to interpret “black box” machine learning models is using model-agnostic techniques that infer explanations from the estimated/trained model by treating it as a black box (Ribeiro et al., 2016b). In this application, we rely on the Local Interpretable Model-agnostic Explanations (LIME) technique to interpret the class membership component of the GP-LCCM. LIME (Ribeiro et al., 2016a) learns an interpretable model on top of the original machine learning model with the aim of interpreting individual predictions. First, LIME generates a new dataset by shuffling the



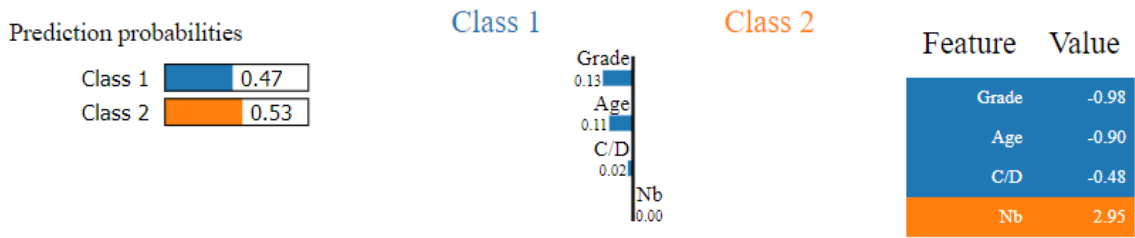
original observations (the socio-economic variables used for clustering). Second, LIME weights the new observations by their closeness to the original dataset. Finally, LIME fits an interpretable model (e.g., linear regression) by using the new shuffled-weighted observations and their associated predictions (class labels) from the original model (GP-LCCM). Figure 6 explains the class predictions of the GP-LCCM with two latent classes. Several individual predictions have been investigated. However, for the sake of brevity, only three observations are presented. The bar charts in Figure 6 portray the importance of each variable, with the value next to each bar being the corresponding weight, while the colors (blue for class 1 and orange for class 2) specify which class the variables contribute to. In the three bar charts of Figure 6, LIME assigns positive and blue weights to Age and Grade while the weights of C/D (ratio of number of cars available over number of licensed drivers per household) and Nb (number of people who are usually present in the car during the trip from home to AUB) are close to zero. This implies that the first class (blue color) is characterized by higher age and grade values which is in line with the corresponding positive parameters from the LCCM and GM-LCCM (Table 16). Moreover, the order of magnitude of the weights, which represent the importance of each variable, is similar to the order of parameter magnitudes from the LCCM and GM-LCCM. Both models from Table 16 have high parameter/mean estimates for Age and Grade while the ones related to C/D and Nb are lower and insignificant according to the LCCM (high p-values). Finally, the first two individuals (Figures 6.a and 6.b) are more likely to belong to the second class (class probabilities higher than 0.5) while the third individual (Figure 6.c) is more likely to belong to the first class (class probability = 0.75). This shows that although Gaussian Processes make the class components less transparent, local interpretability can still be

achieved by relying on model-agnostic techniques. It is to be noted that local interpretability is sometimes much more relevant for model explainability than abstract global interpretation techniques (Montavon et al., 2018), especially given that the former targets individual explanations that can offer better in-depth realization of features contribution/importance in smaller groups of individuals (Kopitar et al., 2019).

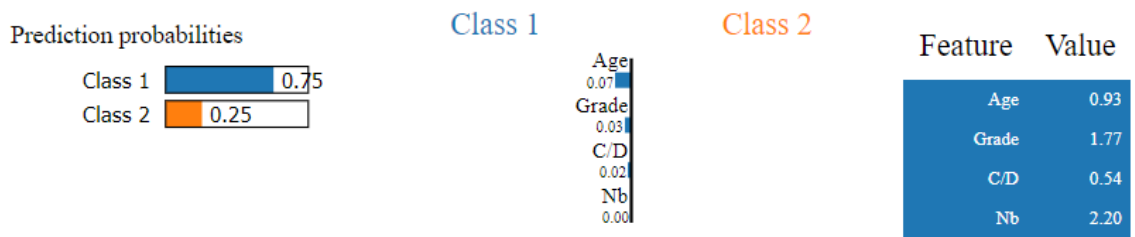
As for the computational times, the LCCM and GM-LCCM took on average less than a minute to converge while the GP-LCCMs with two and three classes took around 5 and 35 minutes, respectively. This difference in runtime is expected due to the nonparametric nature of GPs whose inference requires inverting the covariance matrix with cost of scale  $O(N^3)$ . However, it is to be noted that all runs were performed on a machine with a single core, meaning the implementation was not optimized to take full advantage of modern computational hardware (e.g., GPU) which could make computational overheads less relevant.



a) Individual 1: Age = 24, Grade = 0, C/D = 0.5, Nb = 0



b) Individual 2: Age = 24, Grade = 0, C/D = 0.67, Nb = 5



c) Individual 3: Age = 45, Grade = 16, C/D = 1, Nb = 4

Figure 6: Explaining three individual class predictions of the GP-LCCM with two classes ( $K = 2$ ) using LIME

### 6.3.2. London Case Study

#### 6.3.2.1. Model Specification

The same model specification as in the first trial of section 4.3.1 is assumed. The latent classes of both models, GBM-LCCM and GP-LCCM, are characterized by the available socio-economic variables  $age_n$  (continuous variable representing the age of decision-maker  $n$ ),  $female_n$  (binary variable that equals to 1 if decision-maker  $n$  is female and 0 otherwise),  $car\_own_{n1}$  (binary variable that equals to 1 if the number of

cars in the household of decision-maker  $n$  is more than 0 but less than one per adult and 0 otherwise),  $car\_own_{n2}$  (binary variable that equals to 1 if the number of cars in the household of decision-maker  $n$  is one or more per adult and 0 otherwise), and  $license_n$  (binary variable that equals to 1 if decision-maker  $n$  has a driving license and 0 otherwise). The class-specific choice models are characterized by alternative-specific travel time and travel cost coefficients in addition to a constant in the utility of the car alternative. The class membership component of the GBM-LCCM is characterized by a full covariance structure. As for the kernel function of the GP-LCCM, a Matérn kernel with a smoothness parameter ( $\nu$ ) of 1.5 is selected since, compared to other kernel functions, it generated better in-sample goodness-of-fit measures, better out-of-sample prediction accuracy, and reasonable parameter estimate signs and magnitudes.

#### 6.3.2.2. Estimation Results

Summary statistics of the GBM-LCCM and GP-LCCM are shown in Table 17. The LCCM is not considered since it was shown in Chapter 4 (Section 4.3.1.1) to generate positive cost coefficients. As discussed in Chapter 4, the GBM-LCCM was able to identify up to four latent classes. The proposed GP-LCCM was also able to identify four latent classes. However, the GP-LCCM models with three and four classes generated positive and significant travel cost coefficients and as such are ignored since models with counter-intuitive cost coefficients cannot be used for meaningful predictions and policy analysis. Moreover, the GP-LCCMs with three and four classes improved the joint LL drastically (greater than -270). This vast improvement in the joint LL (more than 87%) of these two models is a clear case of overfitting.

Comparing the models with two latent classes, the GP-LCCM improves the in-sample goodness-of-fit (LL) and the out-of-sample prediction accuracy (Pred. LL) by

7.6% and 8.2%, respectively. The two-class GP-LCCM also outperforms the three-class GBM-LCCM in terms of marginal choice log-likelihood and prediction accuracy by 3.8% and 2.1%, respectively. Comparing the two-class GP-LCCM and the four-class GBM-LCCM, the former shows slightly better marginal choice log-likelihood while the later exhibits slightly better prediction accuracy.

Table 18 presents the class-specific choice estimates of the GBM-LCCM and GP-LCCM with two latent classes (values between parentheses are p-values) as well as the corresponding VOTs. Note that we are presenting the results of the two-class GBM-LCCM and not the four-class model for consistency of comparison with the two-class GP-LCCM. All travel time and travel cost parameter estimates have the expected negative sign. According to the GBM-LCCM, individuals from the second class are highly insensitive towards travel cost of PT (p-value = 0.94) and individuals from the first class seem to be slightly insensitive towards travel cost of car (p-value = 0.15). However, according to the GP-LCCM, individuals from both classes seem to be highly insensitive towards travel cost of PT (p-value = 0.99 and 0.98). In addition, both models indicate that the first class is characterized by higher VOT of car than the second class. These values are in line with a previous study by Hillel et al. (Hillel et al., 2019) which showed that the VOT of car is 36.61 £/hr. Note that most of VOTs of PT are not estimated due to the high insignificance of the corresponding travel cost estimates.

Next, the latent classes of the two-class GBM-LCCM are described based on the mean matrix of the class membership component (i.e., the Gaussian-Bernoulli Mixture Model) (Table 19). Note that, for the GBM-LCCM, the continuous variable age is standardized. Therefore, a negative value means the latent class is characterized by

young individuals while a positive value means individuals are older than the average (which is 40 years).

*First Class: Young with low car ownership*

The first class is characterized by young individuals ( $\mu_{age} < 0$ ) from both genders (52.3% are males and 47.7% are females) and who belong to households with no cars (43.6%) or less than one car per adult (51.6%). Individuals from this class are almost equally likely to be licensed (54.7%) or unlicensed drivers (45.3%).

*Second Class: Licensed drivers with high car ownership*

The second class includes individuals above 40 years old ( $\mu_{age} > 0$ ) who are mostly licensed drivers (94.5%) and belong to families with moderate to high car ownership (only 7.6% have no cars). The percentage of males (58.7%) is somewhat higher than females (41.3%).

As previously mentioned in Chapter 4, the aforementioned analysis is a strong indication that the proposed GBM-LCCM guarantees a simple interpretation of the latent classes, although the random utility formulation of the class membership component is replaced by a full mixture model.

Next, similarly to the previous application (Section 6.3.1), the two latent classes of the GP-LCCM are locally interpreted using the model-agnostic technique LIME. Figure 7 explains the class predictions of two individuals. The bar charts represent the contribution of each variable to the class prediction and the values next to the bars denote the corresponding weights. The blue and orange colors represent the first and second class, respectively. The second class seems to be characterized by licensed drivers who belong to households with high car ownership per adult since LIME assigns orange and positive weights to the variables `car_own2`, `car_own1` and `license`.

Individuals from the second class are likely to be males above 40 years old since Lime associates the variable age with an orange color and the variable female with a blue color. However, the contribution of age and female variables to the predictions seems to be limited due to the corresponding low weight values. This class interpretation is consistent, to high extent, with the characteristics of the classes of GBM-LCCM. The first individual (Figure 7.a) is a 32-year-old unlicensed female driver from a household with no cars. This individual belongs to the first class with a probability of 98%. The second individual is a 61-year-old licensed male driver who lives in a household with high car ownership ( $\text{car\_own2} = 1$ ). This individual can be assigned to the second class with a probability of 85%.

Table 17: Summary results of the London application

K	Model	Joint LL <sup>a</sup>	LL <sup>b</sup>	Variance <sup>c</sup>	AIC	BIC	Pred. LL
2	<b>GBM-LCCM</b> <b>Full Covariance</b>	-18,660.22	-2,920.92	0	5,887.84	6,048.00	-1,387.89
3		-17,502.93	-2,807.87	0	5,685.74	5,929.47	-1,300.33
4		-17,390.22	-2,703.27	0	5,500.54	5,827.83	<b>-1,262.86</b>
2	<b>GP-LCCM</b>	-2,193.04	<b>-2,700.54</b>	0.03	<b>5,425.08</b>	<b>5,508.64</b>	-1,273.24

a: joint log-likelihood of the GBM-LCCM (Equation 21) and GP-LCCM (Equation 44)

b: marginal choice log-likelihood of the GBM-LCCM (Equation 31) and the GP-LCCM (Equation 49)

c: variance of the marginal choice log-likelihood (LL)

Table 18: Class-specific choice estimates of GBM-LCCM and GP-LCCM (K = 2)

Parameter	GBM-LCCM		GP-LCCM	
	Class 1	Class 2	Class 1	Class 2
<b>ASC (Car)</b>	-0.00350 (0.99)	1.83 (0.00)	-0.672 (0.98)	2.32 (0.00)
<b>Travel Time (PT)</b>	-0.0879 (0.00)	-0.0831 (0.00)	-0.0914 (0.00)	-0.0916 (0.00)
<b>Travel Time (Car)</b>	-0.381 (0.00)	-0.121 (0.00)	-0.258 (0.00)	-0.130 (0.00)
<b>Cost (PT)</b>	-0.428 (0.00)	-0.00270 (0.94)	-0.119 (0.99)	-0.0868 (0.98)
<b>Cost (Car)</b>	-0.211 (0.15)	-0.196 (0.00)	-0.207 (0.04)	-0.210 (0.00)
<b>VOT (£/hr)</b>				
<b>PT</b>	12.32	-	-	-
<b>Car</b>	108.34	37.04	72.22	37.10

Values within parentheses are p-values.

Travel Time variables are in minutes.

Cost variables are in Pound Sterling (£ gbp).

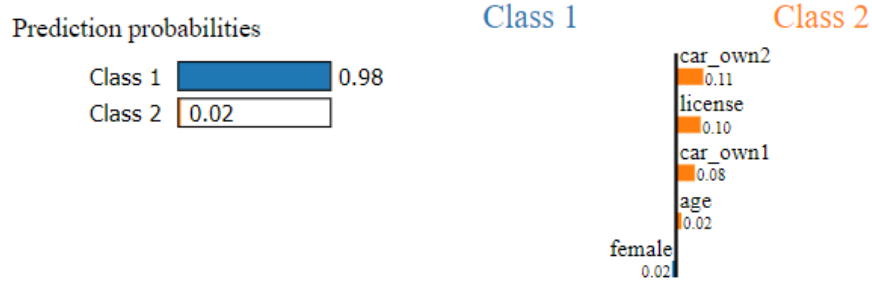
Table 19: Mean matrix of the class membership model of GBM-LCCM (K = 2)

Parameter		Class 1	Class 2
<b>age**</b>	Continuous	-0.261	0.264
<b>female</b>	Yes	0.523	0.413
	No*	0.477	0.587
<b>license</b>	Yes	0.547	0.959
	No*	0.453	0.041
<b>car_own<sub>0</sub></b>	0*	0.436	0.076
<b>car_own<sub>1</sub></b>	] 0 – 1 [	0.516	0.495
<b>car_own<sub>2</sub></b>	≥ 1	0.048	0.429

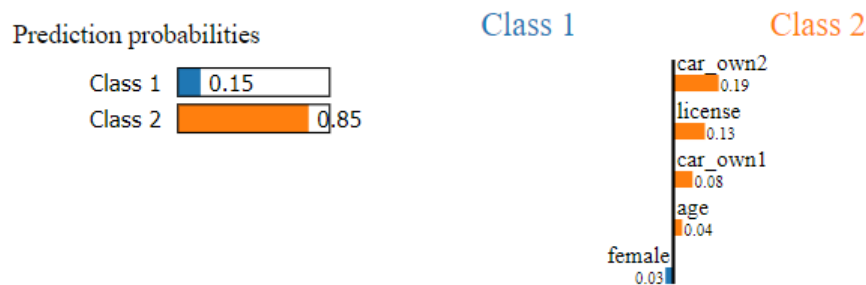
\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1





a) Individual 1: age = 32, female = 1, license = 0, car\_own<sub>1</sub> = 0, car\_own<sub>2</sub> = 0



b) Individual 2: age = 61, female = 0, license = 1, car\_own<sub>1</sub> = 0, car\_own<sub>2</sub> = 1

Figure 7: Explaining two individual class predictions of the GP-LCCM with two classes ( $K = 2$ ) using LIME

### 6.3.3. *Swissmetro Case Study*

#### 6.3.3.1. Model Specification

In this application, we only compare GP-LCCM to LCCM. GBM-LCCM is not considered since no continuous variables are used for clustering. The latent classes of the two models are characterized by the categorical variables AGE, MALE, INCOME, FIRST, LUGGAGE and PURPOSE as shown in Table 20. As for the class-specific choice models, the utilities of the three alternatives are specified using generic travel time and travel cost coefficients in addition to alternative-specific constants for the Train and Car alternatives. We make use of the L-BFGS-B optimizer (C. Zhu et al., 1997) to constrain the signs of the travel time and travel cost parameters since both models generated counter-intuitive positive signs for travel cost and/or travel time

coefficients when using the unbounded optimizer BFGS (Nocedal & Wright, 2006).

The number of latent classes is varied from 2 to 10 and the models are estimated 5 times with different random initializations to assess the stability of the models.

Table 20: Variables used to define the latent classes

<b>Variable</b>	<b>Description</b>	<b>Levels</b>
<b>AGE</b>	The age class of respondents	Age $\leq$ 24*; 24 < Age $\leq$ 39; 39 < Age $\leq$ 54; 54 < Age $\leq$ 65; Age > 65
<b>MALE</b>	The respondent's gender	1: Male; 0: Female
<b>INCOME</b>	The respondent's income per thousand CHF per year	INCOME < 50*; 50 $\leq$ INCOME $\leq$ 100; INCOME > 100; M_INCOME: unknown income
<b>FIRST</b>	First class traveler	0: no; 1: yes
<b>LUGGAGE</b>	Number of luggage the respondent carries during a trip	0: none; 1: one piece; 2: more than one piece*
<b>PURPOSE</b>	Purpose of the trip	1: Commuter; 2: Shopping; 3: Business; 4: Leisure*

\*: level kept as a base

### 6.3.3.2. Estimation Results

Table 21 shows the summary statistics of LCCMs and in particular the Log-Likelihood (LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood of the test sample (Pred. LL). Increasing the number of latent classes beyond 5 for LCCM resulted in some travel time and travel cost parameters with a zero value while other parameter estimates from both sub-components (i.e., the class membership model and the class-specific choice model) had very high standard errors. It is clear that 5 is the optimal number of classes for LCCM since the corresponding model has the lowest LL, AIC, BIC and Pred. LL.

Table 21: Summary results of LCCM

<b>K</b>	<b>Nb of Parameters</b>	<b>LL</b>	<b>AIC</b>	<b>BIC</b>	<b>Pred. LL</b>
2	23	-5,930.76	11,907.52	12,069.75	-1,490.62
3	42	-5,202.71	10,489.41	10,785.67	-1,329.94
4	61	-4,870.51	9,863.02	10,293.29	-1,245.18
5	80	-4,687.99	9,535.99	10,100.28	-1,233.69

Table 22 presents the same measures shown in Table 21 in addition to the joint LL of the GP-LCCMs. Estimating models with 8 or more classes generated zero values for some of the constrained parameters (travel time and/or travel cost). Similarly to the previous case studies, a manual search was conducted to find the optimal kernel function or combination of kernels. Consequently, a Matérn kernel with a smoothness parameter ( $\nu$ ) of 1.5 was used for all GP-LCCMs. Results show that the 7-class model has the lowest joint LL, LL, AIC, BIC and Pred. LL. Compared to the GP-LCCMs with the same number of classes, the LCCMs have better LL (for  $K = 4$  and  $5$ ) and better prediction accuracy (for  $K = 2, 3, 4$  and  $5$ ). However, the optimal GP-LCCM with 7 classes outperforms the optimal LCCM with 5 classes over all statistical measures. Results show that the proposed GP-LCCM has the ability to improve the representation of unobserved heterogeneity by identifying a higher number of latent classes, thus improving the model fit and generalization performance. Compared to the best LCCM ( $K = 5$ ), the best GP-LCCM ( $K = 7$ ) improves the in-sample goodness-of-fit (LL) and the prediction accuracy (Pred. LL) by 1% and 2%, respectively. Finally, the 7-class GP-LCCM generates VOTs between 0.01 and 8.96 CHF/min. In a previous study by Han (2019), a nonlinear-LCCM with 6 classes generated VOTs between 0.03 and 6.86 CHF/min while Bierlaire et al. (2001) showed, using different MNL and Nested Logit specifications, that the VOT is around 1.2 CHF/min. Given that a relatively high

number of classes (seven) is estimated, it is expected to have few classes with high or low Values of Time due to the insignificance of some parameter estimates of travel time and/or travel cost. Details of the best LCCM (K = 5) and best GP-LCCM (K = 7) are presented in Appendix C.

Finally, all LCCMs (K = 2 to 5) took less than a minute to converge, while the computational time of the GP-LCCMs varied between 5 and 55 minutes. As previously mentioned, the implementation was not optimized to take full advantage of modern computational hardware which could make computational overheads less relevant.

Table 22: Summary results of GP-LCCM

<b>K</b>	<b>Nb of Parameters</b>	<b>Joint LL<sup>a</sup></b>	<b>LL<sup>b</sup></b>	<b>AIC</b>	<b>BIC</b>	<b>Pred. LL</b>
2	10	-5,930.02	-5,916.43	11,852.86	11,923.39	-1,493.52
3	18	-4,879.78	-5,176.06	10,388.11	10,515.08	-1,354.44
4	24	-4,260.21	-4,878.84	9,805.68	9,974.97	-1,263.21
5	30	-3,872.87	-4,825.55	9,711.11	9,922.72	-1,256.62
6	36	-3,564.44	-4,742.13	9,556.26	9,810.19	-1,237.09
<b>7</b>	<b>42</b>	<b>-3,346.24</b>	<b>-4,649.20</b>	<b>9,382.39</b>	<b>9,678.65</b>	<b>-1,213.44</b>

a: joint log-likelihood of the GP-LCCM (Equation 44)

b: marginal choice log-likelihood of the GP-LCCM (Equation 49)

#### 6.4. Conclusion

This chapter applied the proposed GP-LCCM to three different mode choice applications and compared it to the GBM-LCCM and/or traditional LCCM. The findings of two case studies (AUB and Swissmetro) indicate that the GP-LCCM allows for a higher degree of flexibility by estimating more latent classes than the benchmark models. Moreover, it is capable of improving the in-sample goodness-of-fit measures and the out-of-sample predictive power by up to 7.6% and 8.8%, respectively. This is due to the fact that GPs rely on a nonparametric structure that lessens the restrictive

parametric assumptions of GBM-LCCM and allows more flexibility than the linear specifications of the class membership utilities of traditional LCCMs. As for the London case study, the proposed GP-LCCM generated the same number of classes as the proposed GBM-LCCM. However, GP-LCCMs with more than two classes were prone to overfitting. This may be due to the fact that nonparametric models are capable of achieving a high degree of flexibility that may eventually result in overfitting (Nisbet et al., 2017). Nevertheless, the GP-LCCM with just two latent classes outperformed the GBM-LCCM with two and three classes and resulted in similar goodness-of-fit and generalization measures as the GBM-LCCM with four classes. Results of the three applications also showed that the use of Gaussian Processes did not compromise the economic and behavioral interpretation of the class-specific choice models. In fact, marginal effects and economic indicators such as VOTs can be easily derived from the model.

Three limitations can be identified. First, the interpretation of the latent classes becomes less transparent. However, latent classes can still be interpreted locally by means of model-agnostic techniques such as LIME. Second, the use of GPs places additional burden on the modeler to select an appropriate kernel function or a combination of kernel functions. Future work could explore ways to automate this task by automatically searching for the kernel structure that would maximize the marginal choice log-likelihood of the overall model. Third, the nonparametric nature of GPs make the estimation process computationally expensive, especially for large datasets. Such limitation could be overcome by using Sparse Gaussian Processes that significantly reduce the time complexity (Titsias, 2009). Finally, although three different mode choice applications have been considered in this chapter, the proposed

model should be applied to different type of datasets to examine whether the same findings could be reached or not.

## CHAPTER 7

### POLICY ANALYSIS

In this chapter, we compare the forecasts of different policies given by the traditional LCCM and the two proposed models, GBM-LCCM and GP-LCCM. We consider the case study of the American University of Beirut to compare the market shares predicted by each of the three models. Section 7.1 discusses the forecasting capabilities of theory- and data-driven methods. Section 7.2 describes the forecasting application. Section 7.3 presents the corresponding results. Section 7.4 elaborates on the issue of calibrating the constants. Section 7.5 concludes.

#### **7.1. Discrete Choice Models and Machine Learning for Policy Analysis**

Discrete choice models are usually used for purposes of forecasting and policy analysis to predict behavior of decision-makers in counterfactual settings and answer what-if questions. Counterfactual settings include, but are not limited to, changes in the choice set (e.g., existing alternatives become unavailable, new alternatives become available), changes in some attributes of the available alternatives, changes in some characteristics of decision-makers, new decision-makers, etc. (Manski, 2013). This is possible with discrete choice models since they are based on random utility theory that defines a decision-maker's utility as a function of his/her characteristics and attributes of the available alternatives. With such approach, a researcher can easily predict a decision-maker's choices in new settings by simply changing the corresponding values of his/her characteristics and/or alternatives' attributes assuming stability of preferences (i.e., of the values of parameter estimates) over time. A transportation researcher can

then answer questions such as “what if we raise taxes on cars or decrease public transport fares?”.

Any model used for policy analysis must be interpretable and provide meaningful extrapolations (Aboutaleb et al., 2021; Manski, 2013). Figure 8 (Aboutaleb et al., 2021) shows the demand  $y$  for a product or service as a function of its price  $X$ . The target is to build a model that can estimate the demand as a function of price,  $P(y|X)$ , then use this model to predict the demand in case of price changes and, more importantly, to extrapolate the demand  $y$  for values of  $X$  beyond the range of historically observed prices. A theory is most needed to extrapolate beyond the range of observed values (Varian, 1993). Econometric models such as discrete choice models are theory-driven models that rely on statistical assumptions to enable meaningful extrapolations. For instance, the researcher’s a priori assumption regarding the problem presented in Figure 8 is that an increase in price  $X$  should affect the demand  $y$  negatively. The estimated linear econometric model with a negative slope (blue trend) confirms the researcher’s a priori assumption and, consequently, the model can be used for meaningful extrapolations. Similarly for mode choice modeling, a model should generate reasonable parameter estimates that can be used for meaningful forecasts. A model with counter-intuitive parameter estimate signs (e.g., positive travel cost coefficient) cannot be used for policy analysis even if the model guarantees high goodness of fit.

On the other hand, machine learning models are mostly data-driven methods that target maximizing fit and prediction accuracy. Regarding the problem presented in Figure 8, a supervised machine learning model, most probably of the second polynomial order, is selected to model the non-linearity of the data. The fitted model (red trend)



predicts perfectly the demand with respect to price changes as long as the changes are within the range of historically observed prices. This stems from the fact that models that only consider maximizing fit without relying on a priori theory do not guarantee meaningful extrapolations (Aboutaleb et al., 2021).

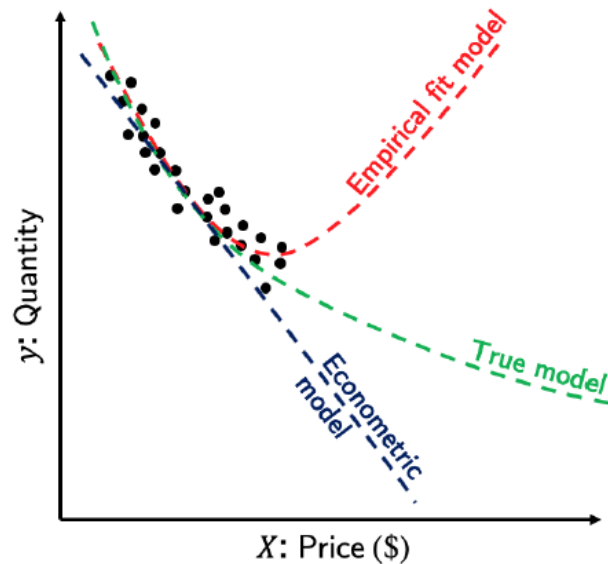


Figure 8: Demand for a product or service as a function of its price (Aboutaleb et al. (2021))

As for the two models that are proposed in this dissertation, the GBM-LCCM and GP-LCCM, we believe that they should provide meaningful forecasts and extrapolations, similarly to the traditional econometric LCCM as we will justify why shortly. The two proposed models are also econometric LCCMs with a more flexible class membership component. The class-specific choice components of both the GBM-LCCM and GP-LCCM are based on random utility theory, similarly to the traditional LCCM, which guarantees meaningful forecasts and extrapolations as long as the class-specific parameter estimates conform to a priori expectations and have the right intuitive signs. In other words, the three models have the same class-specific choice formulation,  $P(y_n|X_n, q_{nk} = 1, \beta_k)$ , while the main difference is the formulation of the latent classes / clusters. Consequently, we hypothesize that the three models should

provide meaningful forecasts even for historically unobserved values (e.g., prices not available in the train dataset), since the class-specific choice components of the three models are based on the same random utility theory.

## **7.2. Application**

A policy analysis application is developed using the case study of the American University of Beirut to compare the forecasting performance of the two proposed models and the traditional LCCM. The sample enumeration method, which averages the choice probabilities for a sample of individuals, is used to forecast the demand (i.e., aggregate predictions) for the proposed services, Shared Taxi (ST) and Shuttle (SH). The utilities and probabilities of all alternatives are estimated using the three models developed in Chapters 4 and 6 (LCCM, GBM-LCCM, and GP-LCCM), and expansion factors<sup>9</sup> are used to weigh up to the total population. We define a base case scenario for the two potential services (shared-taxi and shuttle). This scenario is presented in Table 23 and includes the defined fare of a one-way trip by shared-taxi, fare of a one-way trip by shuttle, travel time by shared-taxi, access and in-shuttle travel times of shuttle, and headway (inversely related to shuttle frequency) between shuttles. For the shuttle service, we assume a reasonable location near AUB for the satellite parking. The travel time and cost of commuting by car are those reported by the respondents in the survey. For the other modes, travel time and cost vary by region of residence (A, B, C). For more details about the base case scenario, readers can refer to Sfeir et al. (2020).

---

<sup>9</sup> For students and faculty samples: gender distribution is used. For staff sample: grade distributions is used.

Table 23: Attributes of the shared-taxi and shuttle options (adapted from Sfeir et al. (2020))

Attributes	Shared-Taxi	Shuttle
One-Way Fare (L.L.)	4,000 (A) 5,000 (B) 8,000 (C)	3,000
Travel Time (Shared-Taxi)	1.3*T <sub>C</sub> (A) 1.25*T <sub>C</sub> (B) 1.2*T <sub>C</sub> (C)	-
Access Travel Time (Shuttle)	-	1*T <sub>C</sub> (A) 0.9*T <sub>C</sub> (B) 0.85*T <sub>C</sub> (C)
In-Shuttle Travel Time (minutes)	-	15
Shuttle Headway (minutes)	-	20

A: Region A is within 5 km from AUB

B: Region B is 5 km to 10 km away from AUB

C: Region C is 10 km or more away from AUB but within Greater Beirut Area (GBA)

T<sub>C</sub>: Travel time by car as reported by the respondents in the survey

### 7.3. Results

In this section, we present the forecasts of different models under the base case scenario (Section 7.3.1). The five models that were developed and presented in Section 6.3.1 are considered. Namely, the models are: LCCM (K = 2), GM-LCCM (K = 2), GM-LCCM (K = 3), GP-LCCM (K = 2), and GP-LCCM (K = 3). Next, we perform a sensitivity analysis to investigate further the forecasting performance of the different models (Section 7.3.2).

#### 7.3.1. Base Case Scenario

Figure 9 shows the variability of individuals' mode choices in a given week under the base case scenario. The five models generate similar forecasts. This might be due to the fact that the class-specific choice components of all the models are based on the same random utility theory, have the same specifications, and generated similar

results in terms of parameter estimate signs and magnitudes (Section 6.3.1). It is apparent that the AUB population is willing to use the proposed services occasionally rather than regularly. Around 41% of individuals would vary their mode of commute in a given week, 42% to 44% would keep only using their cars while only 5% and 10% would be willing to shift completely to the shared-taxi and shuttle services, respectively. Four different multimodality groups can be defined (ST and SH; ST and Car; SH and Car; ST, SH, and Car) with the car-based group (SH and Car) having the highest share (16.5% to 18.6%) among these four groups.

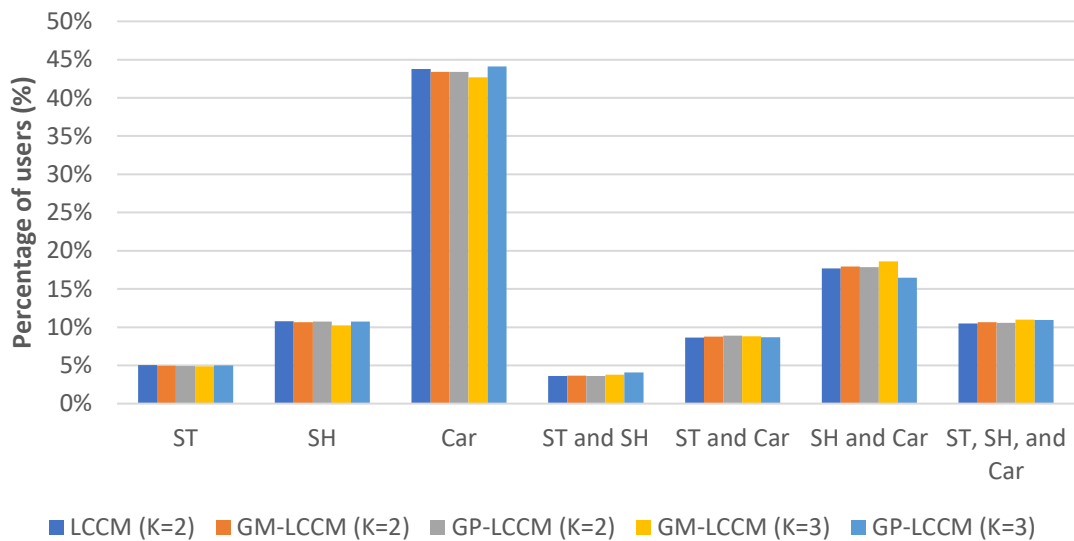


Figure 9: Expected weekly mode share of individuals under the base case scenario

Figure 10 shows the percentage of trips made by each mode in a given week under the base case scenario. Again, the five models produce the same forecasts. Although around 41% of individuals are expected to use multiple modes of travel per week under the base case scenario (Figure 9), only 14% and 25% of the trips are expected to be made by shared-taxi and shuttle, respectively, while 61% would be car trips (Figure 9).

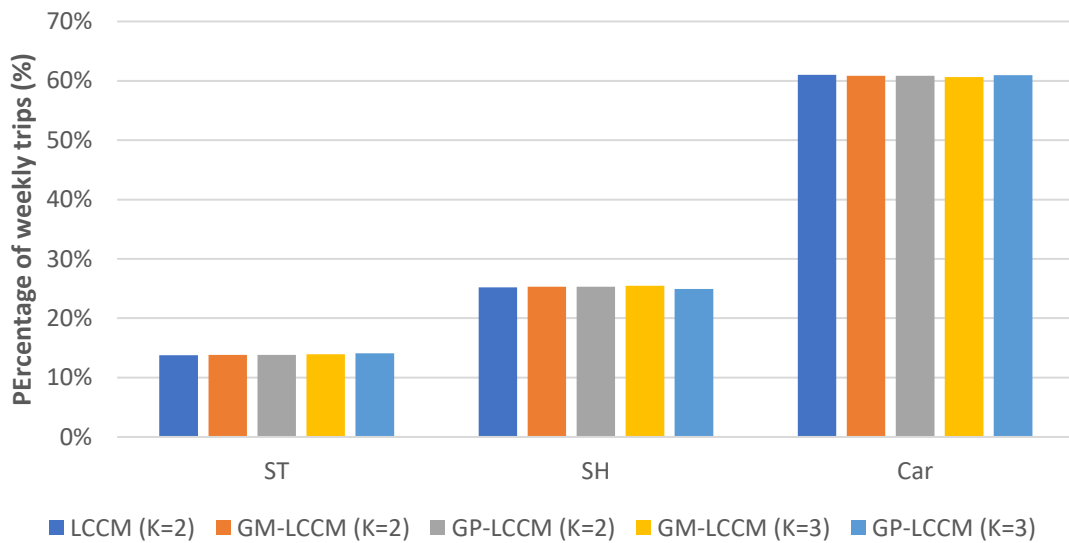


Figure 10: Expected weekly trips per mode under the base case scenario

To better understand the behavioral differences under the base case scenario of the different classes identified, we present the class-specific choices of the LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ). The three-class GP-LCCM is selected since it proved in Chapter 6 to be the best model in terms of in-sample and out-of-sample log-likelihood.

Figures 11 and 12 show respectively the per class variability of individuals' mode or combination of mode choices and per class percentage of trips made by each mode in a given week as predicted by the LCCM with two classes. The first class can be labeled as “unimodal” users who are almost entirely reliant on one mode (Figure 11) in a given week. More than 95% of individuals from the first class are expected to rely on one mode for all their weekly trips to AUB with the majority (68.94%) using only their cars (Figure 11). Overall, around 70% of all the trips from the first class are expected to be made by car (Figure 12). The second class can be labeled as “multimodal” users who would use a combination of different modes in a given week (Figure 11). Although 81.75% of individuals from the second class are expected to have a multimodal style,

almost half of the trips from this class are expected to be car trips (Figure 12). It is to be noted that the first class comprises 53.71% of the population under consideration while the second class comprises the remaining 46.29%.

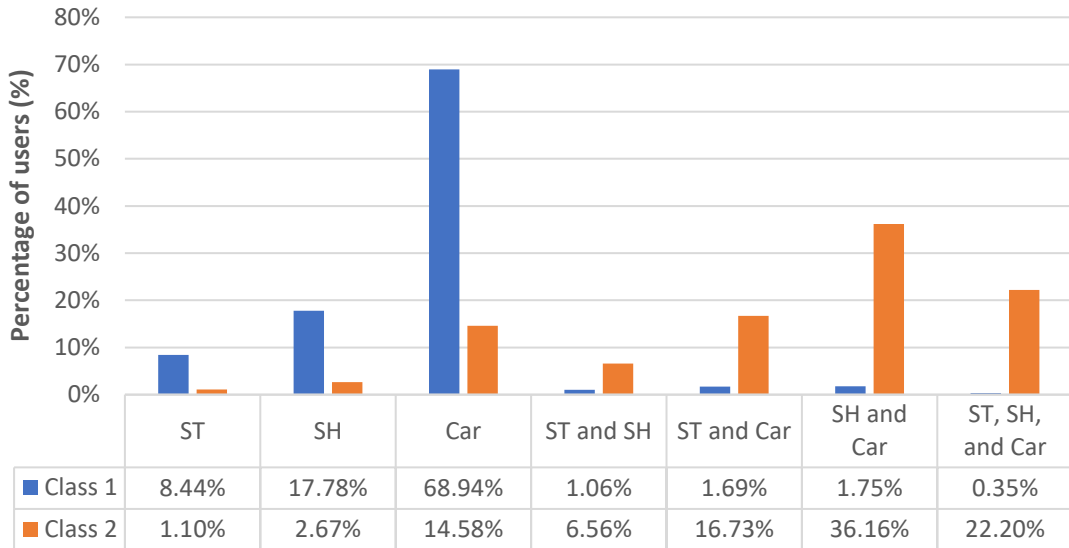


Figure 11: Expected weekly mode share of individuals per class under the base case scenario of LCCM (K = 2)

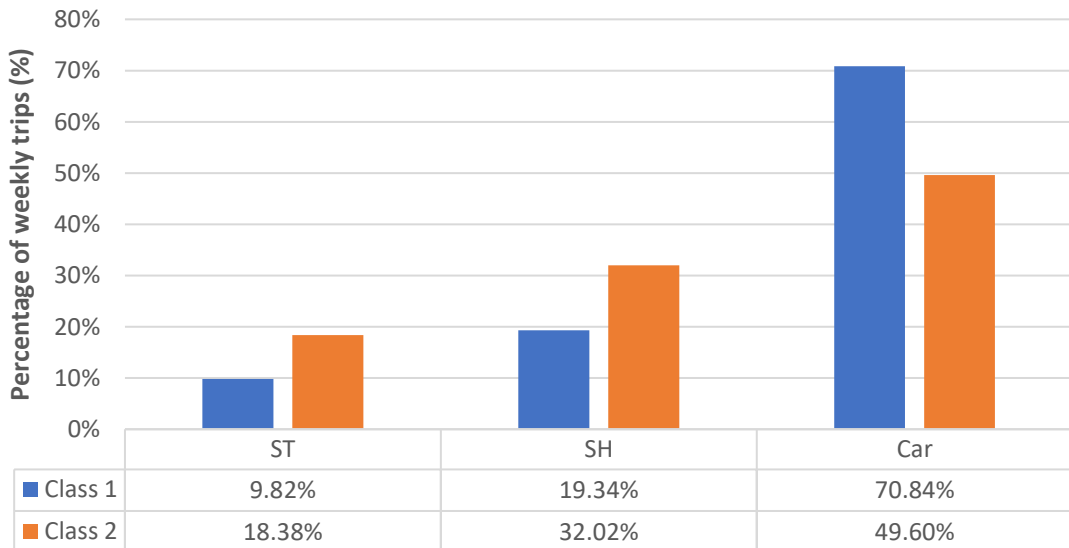


Figure 12: Expected weekly trips per mode and class under the base case scenario of LCCM (K = 2)

Figures 13 and 14 show respectively the per class variability of individuals' mode choices and per class percentage of trips made by each mode in a given week as predicted by the GP-LCCM with three classes. Figure 13 supports the concept of different modality styles (Vij et al., 2013). Similarly to the LCCM, the first class can be labeled as "unimodal" users who are almost entirely reliant on one mode (Figure 13) in a given week. More than 97% of individuals from the first class are expected to rely on one mode for all their weekly trips to AUB with the majority (70.69%) using only their cars (Figure 13). The second and third classes comprise "multimodal" users who would use a combination of different modes in a given week. However, the two "multimodal" classes have different modality compositions. Individuals from the second class are expected to rely more on the two proposed services, shared-taxi and shuttle, than on their own private cars. Around 70% of all the trips from the second class are expected to be made by shared-taxi or shuttle. On the contrary, and although the third class can be also labeled as "multimodal", around 70% of the trips are expected to be made by car (Figure 14). Moreover, Figure 13 shows that individuals from the third class are mainly expected to keep relying on their cars (33.57%) or a combination of their cars and one of the new modes (25.64% belong to the "ST and car" group while 31.10% to the "SH and car" group). Almost no one from the third class will use only the two proposed services in a given week ("ST and SH" group) compared to 14.59% of individuals from the second class who will only rely on the shared-taxi and shuttle services (Figure 13). Moreover, 35.41% of individuals from the second class are expected to use the three different modes in a given week ("ST, SH, and Car" group) compared to only 8.11% of individuals from the third class. Consequently, the third class can be more precisely labeled as "quasi-multimodal with a high reliance on cars". Finally, the first class

comprises 47.91% of the population, the second class 24.36%, and the third class the remaining 27.73%.

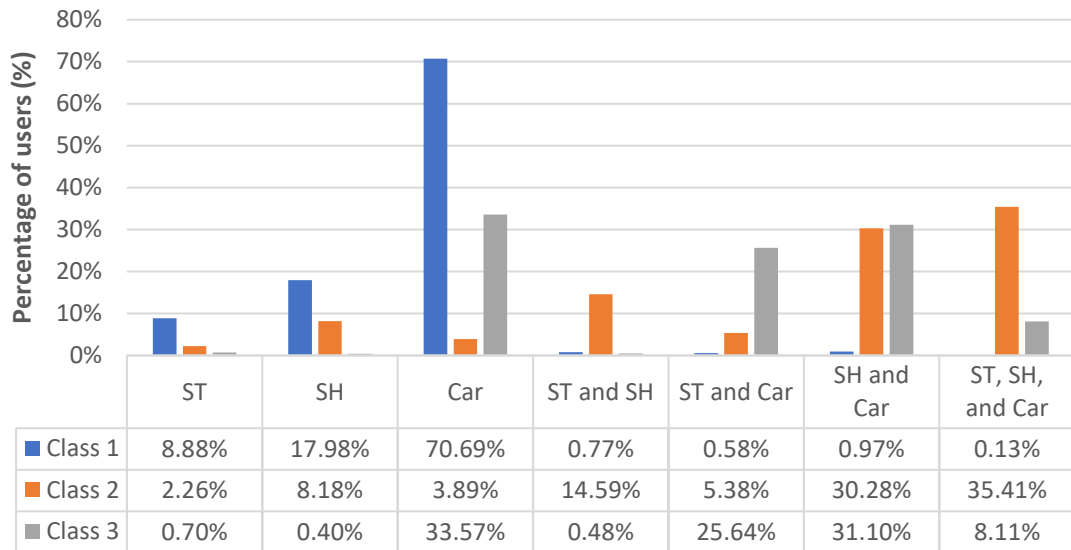


Figure 13: Expected weekly mode share of individuals per class under the base case scenario of GP-LCCM (K = 3)

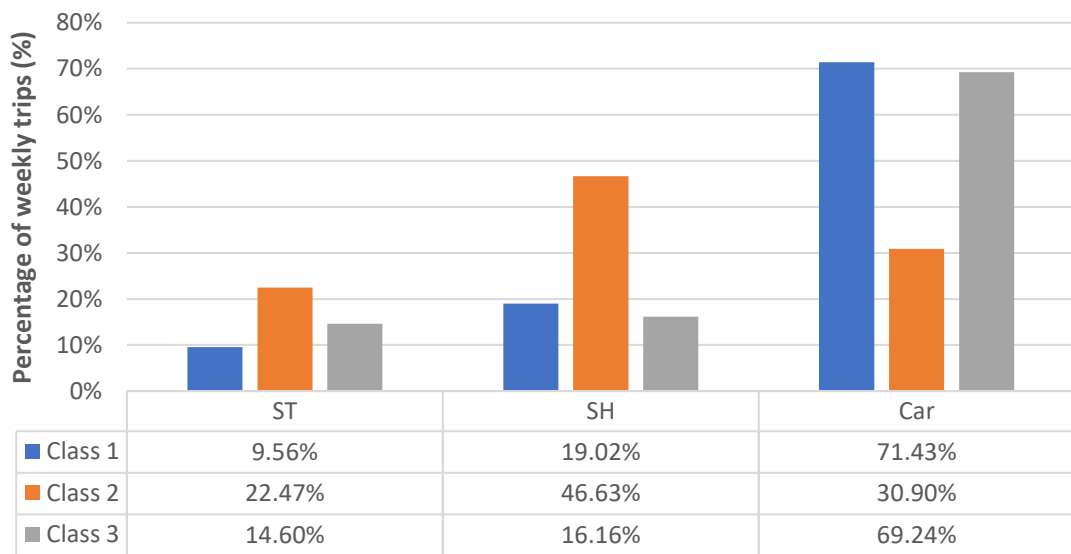


Figure 14: Expected weekly trips per mode and class under the base case scenario of GP-LCCM (K = 3)



### *7.3.2. Sensitivity Analysis*

#### 7.3.2.1. General Sensitivity Analysis

To investigate further the forecasting performance of the models and whether they provide reasonable extrapolations/forecasts, we performed a sensitivity analysis of the demand for the potential services, shared-taxi and shuttle, with respect to their one-way fares. We varied, separately, the one-way fare of each service by multiplying the values of the base case scenario (Table 23) by a factor between 0 and 5. We raised the multiplier factor 0.1 point at a time and after each incremental increase, we predicted the number of weekly trips by each of the two proposed services using the different models (LCCM, GM-LCCM, and GP-LCCM). All models generated similar results, to some extent. For instance, the difference between the number of weekly shared-taxi trips predicted by the GM-LCCM and GP-LCCM with two latent classes and the LCCM did not exceed 3.5%. However, the main and consistent difference was highlighted by the three-class GP-LCCM. Therefore, the percent difference between the forecasts of LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) are presented below.

#### *One-way fare of shared-taxi*

Figures 15 and 16 display, respectively, the percent difference between the weekly shared-taxi and shuttle forecasts of the traditional LCCM ( $K = 2$ ) and the proposed GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shared-taxi. The differences are estimated with respect to the LCCM forecasts. The two figures show exponential trends with opposite directions. For the case of free shared-taxi (multiplier factor = 0), the GP-LCCM predicts more shared-taxi trips than the LCCM by 1.8% (Figure 15) and less shuttle trips than the LCCM by 8% (Figure 16). The GP-LCCM would predict less shared-taxi trips for a multiplier factor higher than 2.8. As for

shuttle trips, the GP-LCCM predicts more trips than the LCCM if the fares are multiplied by a factor higher than 1.3. Finally, for a multiplier factor of 5, the GP-LCCM forecasts 6.7% less shared-taxi trips and around 4% more shuttle trips than the LCCM.

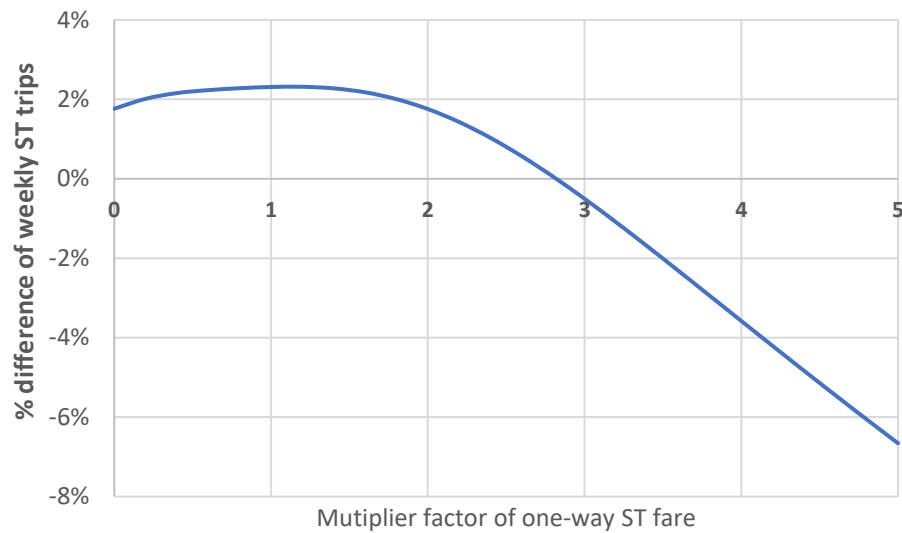


Figure 15: Percent difference of weekly shared-taxi trips as forecasted by LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shared-taxi

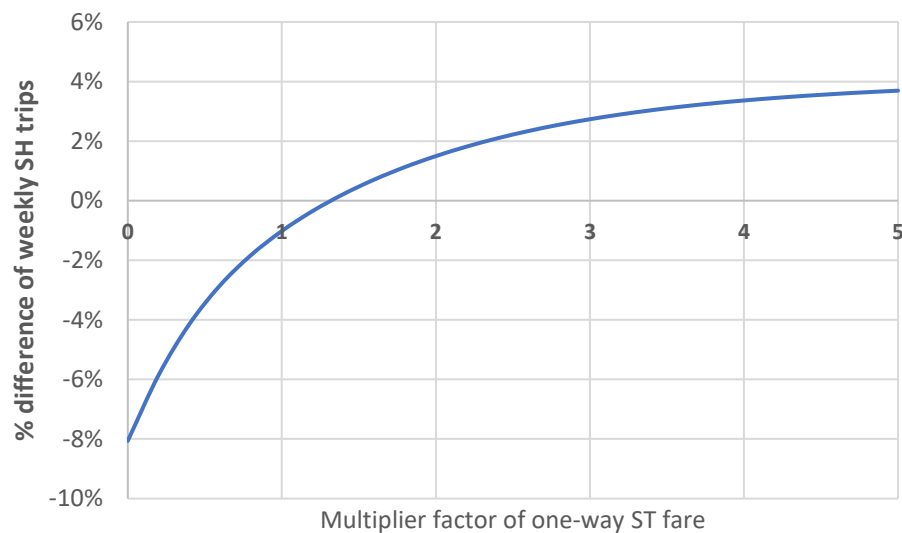


Figure 16: Percent difference of weekly shuttle trips as forecasted by LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shared-taxi

One-way fare of shuttle

Similarly, Figures 17 and 18 display, respectively, the percent difference between the weekly shared-taxi and shuttle forecasts of the LCCM (K = 2) and GP-LCCM (K = 3) with respect to changes in the one-way fare of shuttle. For shuttle fares higher than the base case value (Table 23), the GP-LCCM predicts less shuttle trips (Figure 18) and more shared-taxi trips per week (Figure 17) than the LCCM. On the contrary, for low shuttle fares (multiplier factor less than 1), the GP-LCCM predicts more shuttle trips (Figure 18) and less shared-taxi trips per week (Figure 17). For instance, if shuttle trips between satellite parking hubs and AUB gates are offered for free (multiplier factor = 0), the GP-LCCM would predict 2.8% more shuttle trips and around 6% less shared-taxi trips per week than the LCCM.

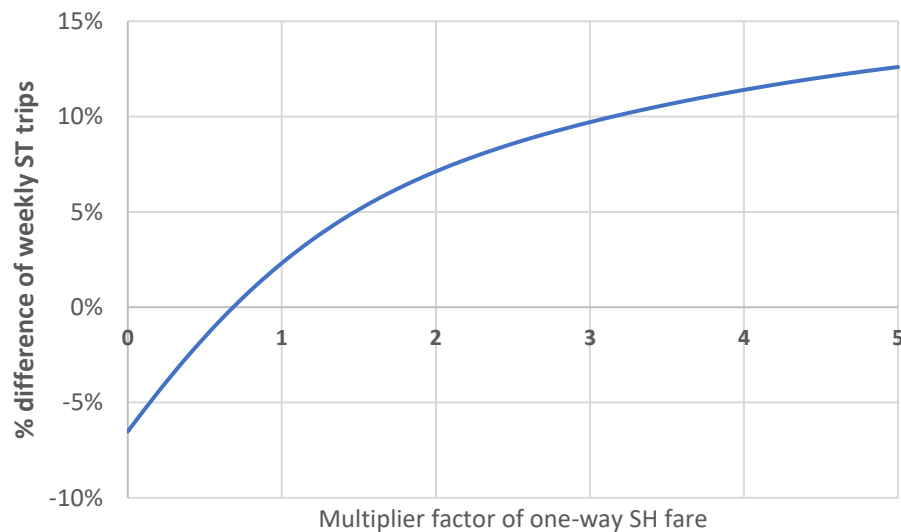


Figure 17: Percent difference of weekly shared-taxi trips as forecasted by LCCM (K = 2) and GP-LCCM (K = 3) with respect to changes in the one-way fare of shuttle

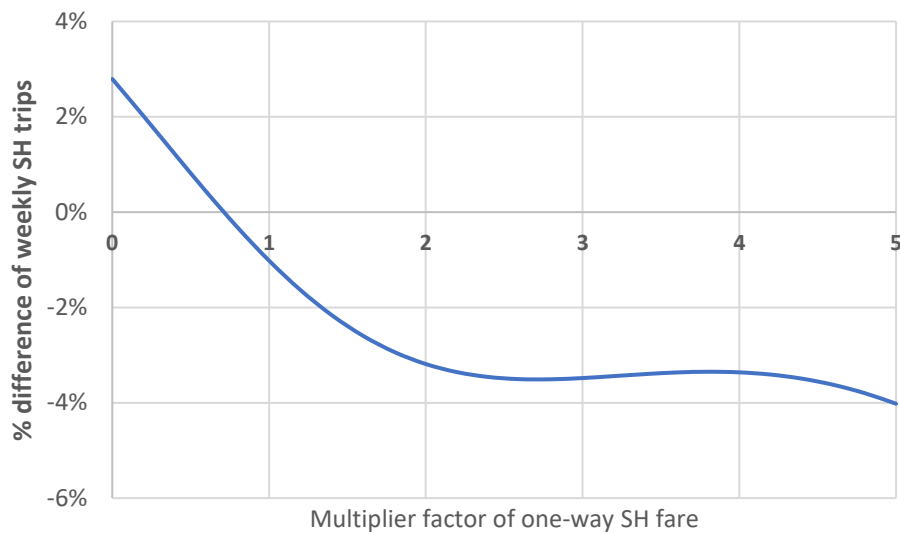


Figure 18: Percent difference of weekly shuttle trips as forecasted by LCCM ( $K = 2$ ) and GP-LCCM ( $K = 3$ ) with respect to changes in the one-way fare of shuttle

It was previously shown in Section 6.3.1 that the GP-LCCM with three classes improves the in-sample and out-of-sample log-likelihood of the LCCM with two classes by 4.5% and 8.8%, respectively (Table 13). While the log-likelihood is the log-sum of the choice probabilities, the forecasts of the sample enumeration method used in the policy analysis application are based on the sum of the choice probabilities. Therefore, it is expected to have similar percent differences between the log-likelihoods and forecasts of the two models. For the base case scenario (multiplier factor of 1), the GP-LCCM predicts more weekly shared-taxi trips than the LCCM by 2.3% and less weekly shuttle trips by 1%. When varying the one-way fare of shared taxi, the percent difference between the forecasts of the two models varies between 2.3% and -6.7% for shared-taxi (Figure 15) and between -8.1% and 3.7% for shuttle (Figure 16). Moreover, changing the one-way fare of shuttle resulted in forecasts that differ between -6.5% and 12.6% for shared-taxi (Figure 17) and between 2.8% and -4% for shuttle (Figure 18).

These statistics are consistent with the in-sample and out-of-sample log-likelihood improvements of the GP-LCCM<sup>10</sup>.

#### 7.3.2.2. Class-Specific Sensitivity Analysis

Sensitivity analysis of the class-specific demand for the potential services with respect to the one-way fare of shared-taxi is also conducted to better understand the behavioral heterogeneity of the different classes. Four scenarios are tested:

- a) Free shared-taxi
- b) 50% discount on the base shared-taxi fare
- c) Base shared-taxi fare (Table 23)
- d) 50% increase of the base shared-taxi fare

Figures 19 and 20 show respectively the per class variability of individuals' mode or combination of mode choices and per class percentage of trips made by each mode in a given week as predicted by the LCCM with two classes under the four above-mentioned scenarios. Changing the one-way fare of shared-taxi does not affect the extent of multimodality of each class. Under the four scenarios, individuals from the first class adopt a weekly unimodal behavior while those belonging to the second class adopt a weekly multimodal behavior (Figure 19). The percentage of ST users from the first class increases from 8% in the base case scenario to around 23% and 53% in scenarios a and b, respectively. Increasing the shared-taxi fare by 50% (scenario d) decreases the ST users from the first class to less than 3% (Figure 19). As for the second class, offering free shared-taxi rides (scenario a), increases the shares of the three multimodal ST-based groups ("ST and SH"; "ST and Car"; "ST, SH and Car") with the

---

<sup>10</sup> Sensitivity analysis with respect to other factors such as travel time and shuttle headway was performed and results (i.e., % difference of forecasted trips) were also consistent with the LL improvements of the GP-LCCM.

“ST and Car” group having the highest increase from around 16% in scenario c to 31% in scenario a (Figure 19). Moreover, the percentage of ST trips from the second class increases from 18% to 41% while the percentage of car trips decreases from 50% to 37% (Figure 20).

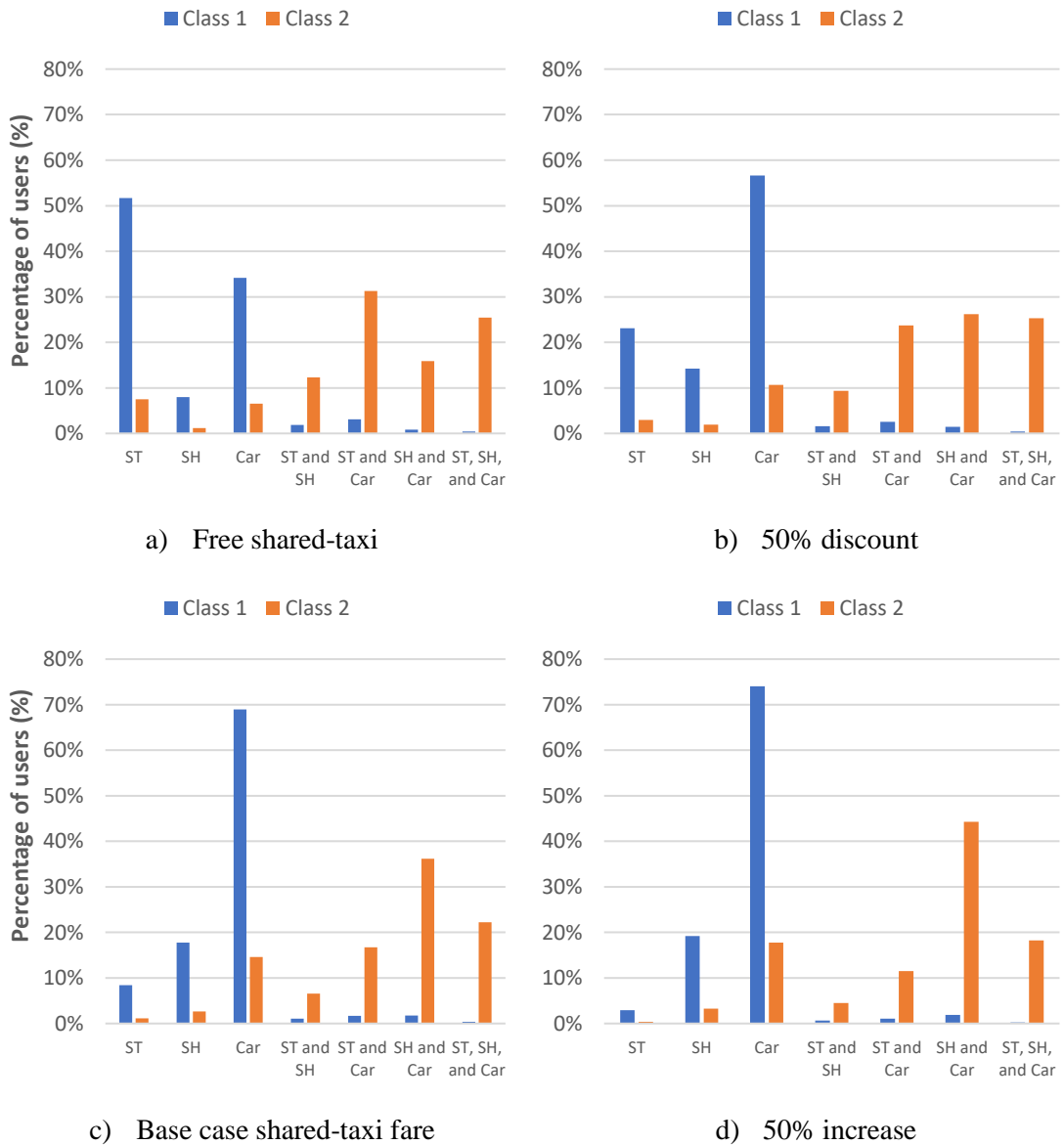


Figure 19: Expected weekly mode(s) share of individuals per class under different one-way shared-taxi fares - LCCM (K = 2)

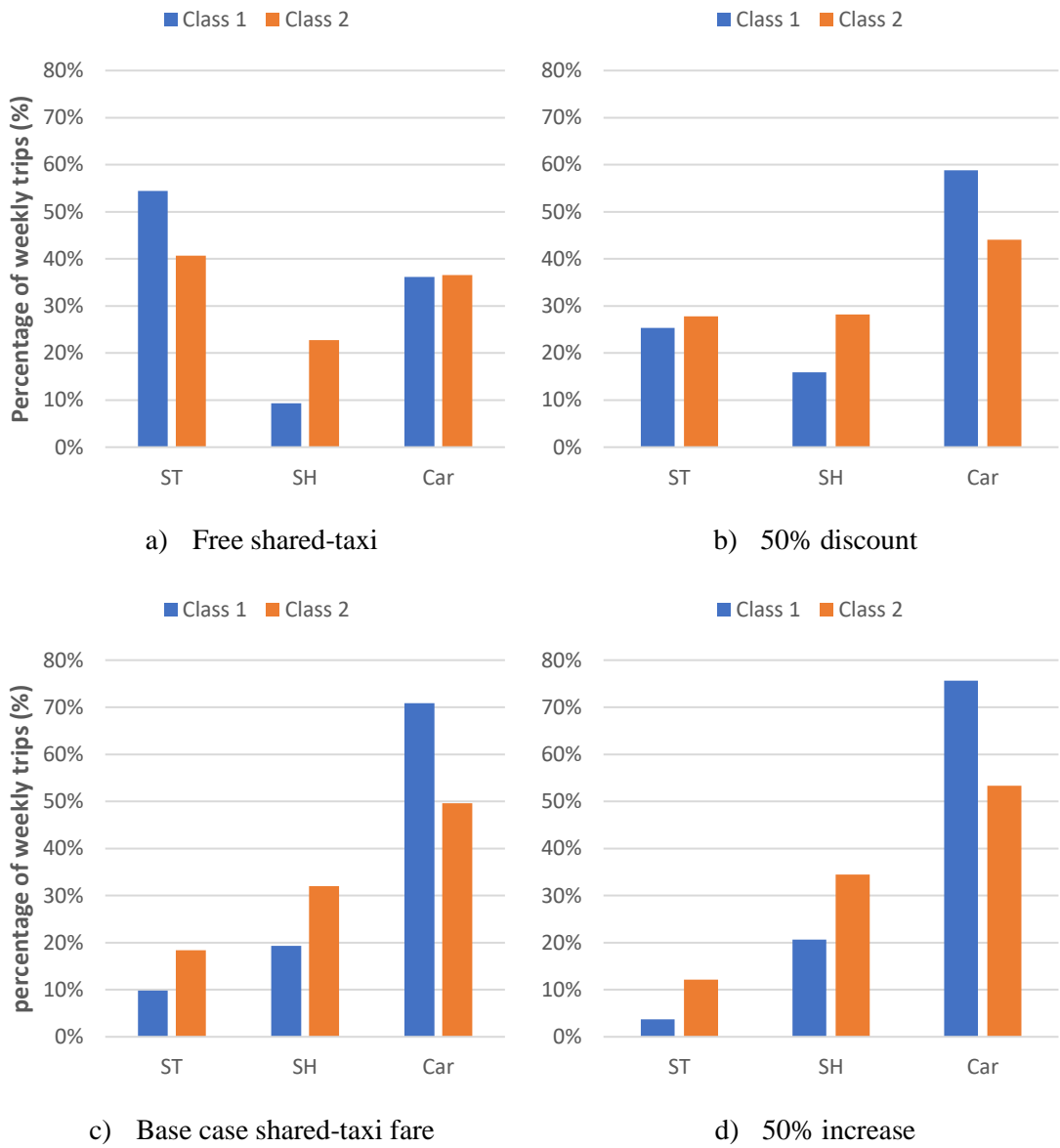


Figure 20: Expected weekly trips per mode and class under different one-way shared-taxi fares – LCCM (K = 2)

Figures 21 and 22 show respectively the per class variability of individuals' mode or combination of mode choices and per class percentage of trips made by each mode in a given week as predicted by the GP-LCCM with three classes under four different scenarios. Similarly to the LCCM, changing the one-way fare of shared-taxi does not affect the extent of multimodality of the classes. As previously mentioned, the first class of the GP-LCCM can be labeled, under the four different scenarios, as “unimodal”, the second class as “multimodal”, and the third class as “quasi-multimodal with a high reliance on cars” (Figure 21). Similarly to the first class of the LCCM, the percentage of ST users from the first class of the GP-LCCM increases from 9% in the base case scenario to around 25% and 55% in scenarios a and b, respectively. Increasing the shared-taxi fare by 50% (scenario d) decreases the ST users from the first class to less than 3% (Figure 21).

As for the second class, offering free shared-taxi rides (scenario a), slightly increases the shares of the three multimodal ST-based groups (“ST and SH”; “ST and Car”; “ST, SH and Car”) with the “ST and SH” group having the highest increase from around 15% in scenario c to 23% in scenario a (Figure 21). However, the percentage of ST trips from the second class increases from 22% to 50% while the percentage of car trips decreases from 31% to 19% (Figure 22). This significant increase in the percentage of shared-taxi trips from the second class, although the changes in the shares of the three multimodal ST-based groups are less significant, can be attributed to the fact that individuals from those ST-based groups are more willing to use the free shared-taxi option than the car and/or shuttle without switching to another group. For instance, someone from the “ST and Car” group who under the base case scenario is willing to



use the shared-taxi twice per week and the car three times per week might switch to using the shared-taxi four times per week and the car once per week.

On the contrary, offering free shared-taxi rides (scenario a) to the third class, increases significantly the share of the “ST and Car” group from 25% to 50% while the shares of the other two ST-based groups (“ST and SH”; “ST, SH, and Car”) remain constant (Figure 21). The percentage of ST trips from the third class increases from around 15% to 35% while the percentage of car trips decreases from 69% to 55% (Figure 22). This high percentage of car trips and the high shares of car-based groups, although the shared-taxi is offered for free, justify the classification of the third class as “quasi-multimodal with high reliance on cars” and the different multimodality behavior between the second and third class.

To sum up, the first class of both the LCCM and GP-LCCM seems to be characterized with the same unimodal behavior. However, multimodal individuals from the second class of the LCCM can be further decomposed in the GP-LCCM into multimodal individuals who are more open to use the new services and multimodal individuals who demonstrate a strong bias towards relying on their private cars. These findings can have important implications for the policies aimed at changing travel behavior since individuals with different behavioral characteristics or modality styles tend to respond differently to such policies.

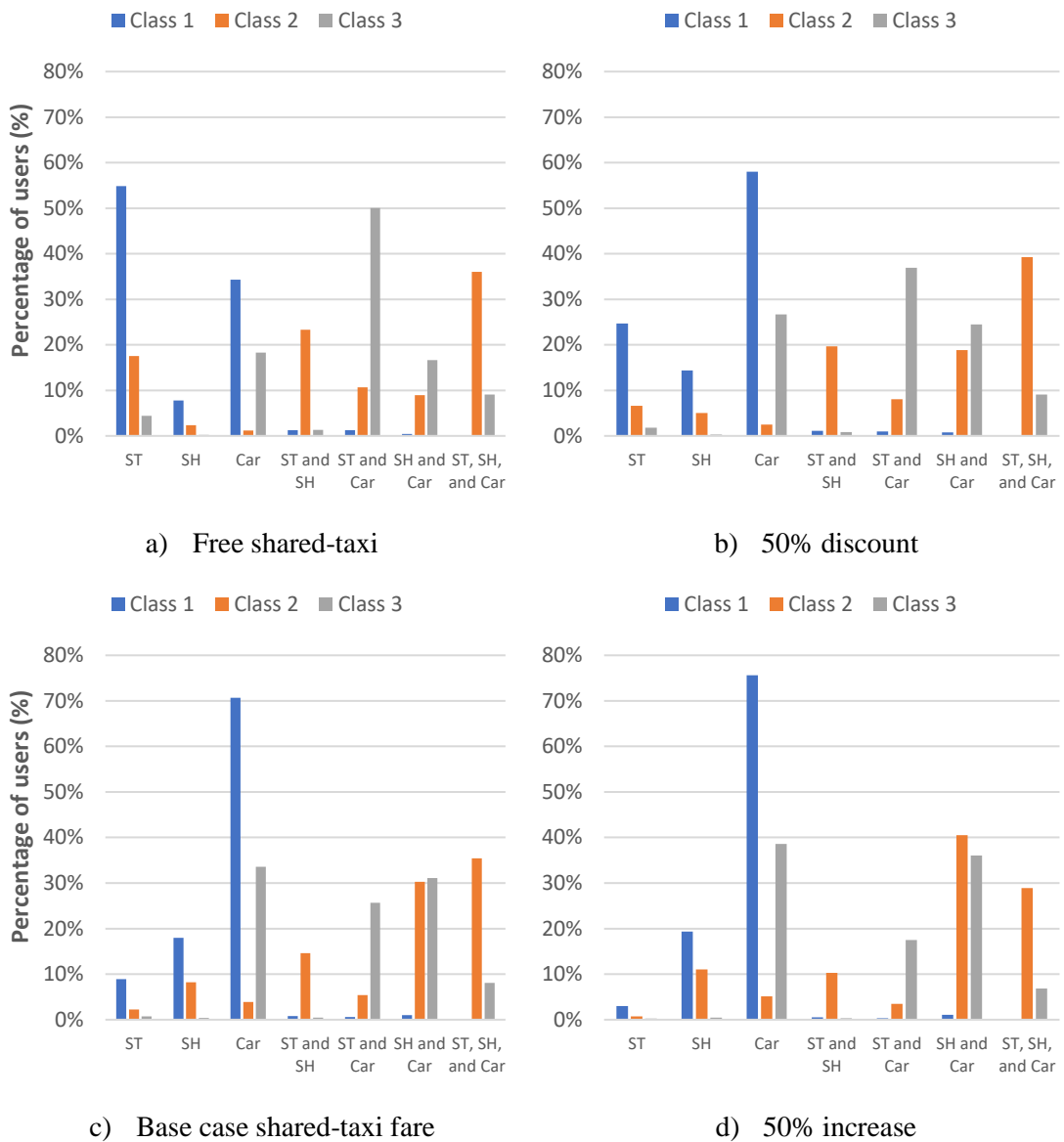


Figure 21: Expected weekly mode(s) share of individuals per class under different one-way shared-taxi fares – GP-LCCM (K = 3)

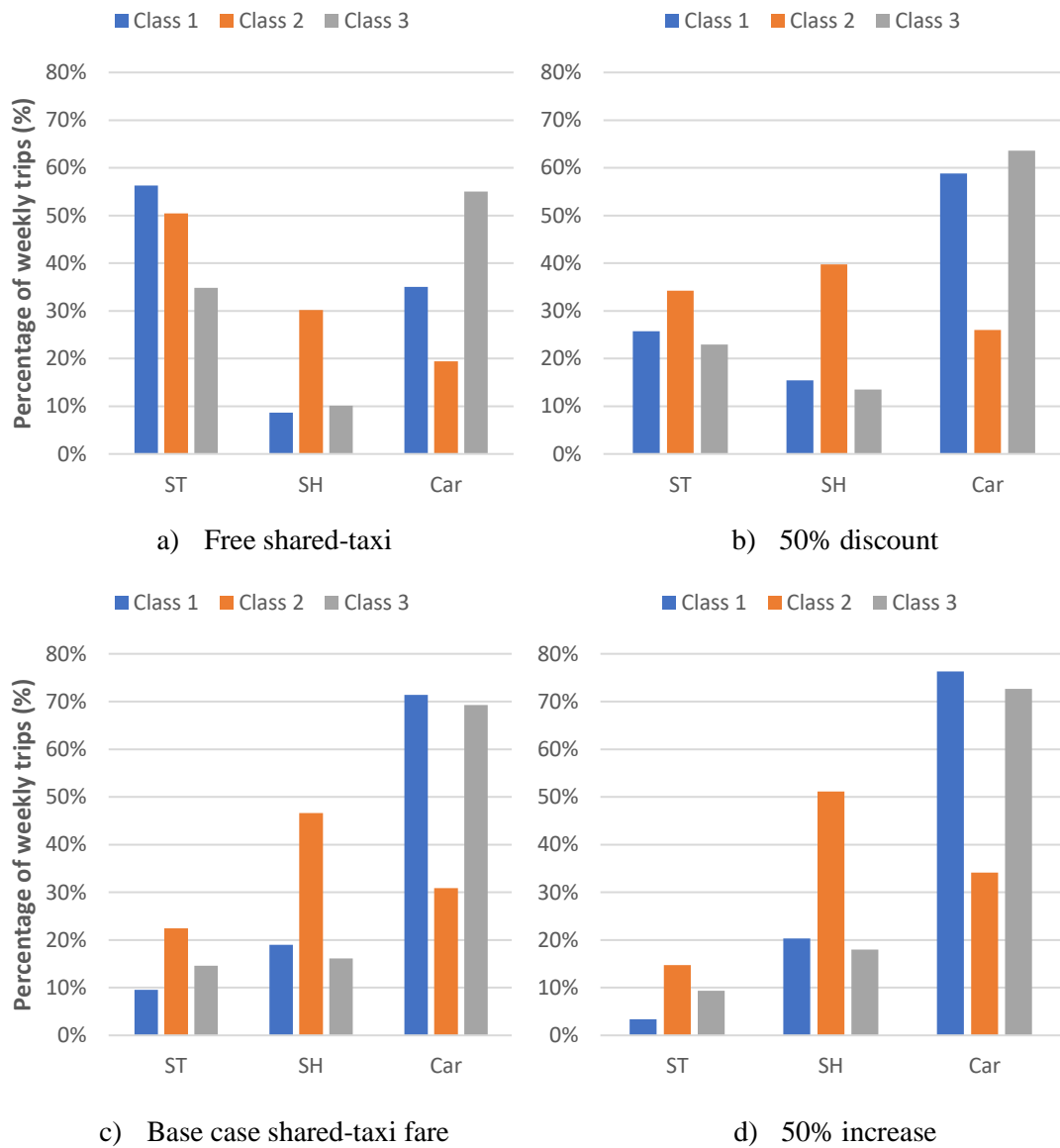


Figure 22: Expected weekly trips per mode and class under different one-way shared-taxi fares – GP-LCCM ( $K = 3$ )

#### 7.4. Calibration of the Constants

The estimated alternative-specific constants in choice models based on stated preferences data must be recalibrated to reflect the real market shares of the forecast area (Cherchi and de Dios Ortúzar, 2006; Glerum et al., 2014; Liu et al., 2019; Train, 2009, to name a few). Usually, real market data is used to correct the constants. Next, the recalibrated model is used to test different policy scenarios and predict the changes in the demand due to changes in some explanatory variables.

However, for this application, we believe the constants cannot be recalibrated. The proposed modes, shared-taxi and shuttle, do not exist in the real life context in Beirut. As for the car option, there are a few studies on the real market shares of the available modes (car and public transport) within the Greater Beirut Area (GBA). However, there is no information on the real weekly frequencies of using these modes. Therefore, recalibrating the constants related to the frequency of using the available modes ( $C_{ST_{h,k}}$ ,  $C_{SH_{i,k}}$  and  $C_{Car_{j,k}}$ ), which are used in the class-specific choice models instead of the traditional alternative-specific constants (Equation 33), would be a very complicated and tricky task.

Nevertheless, as an indication, we compared the forecasting results of the base case scenario to the actual modes reported by the respondents in the survey. We assume that the percentage of respondents who would shift from commuting by car to the proposed services (shared-taxi and shuttle), if they were implemented, should be similar to the percentage of respondents who currently commute by jitney or private taxi in real life. The ratio of respondents who commute by car to those who commute by jitney or private taxi is equal to 2.73<sup>11</sup> while the ratio of respondents who are expected to keep

---

<sup>11</sup> This ratio is estimated from the AUB dataset.

commuting by car only to those who are expected to commute by shared-taxi or shuttle only varies between 2.77 and 2.82 for the five models that were presented in Section 7.3.1. Although this is an indirect validation, we believe it is an indication that the forecasts are within expectations.

## **7.5. Conclusion**

We postulated that the two proposed models should generate meaningful forecasts and extrapolations since the class-specific choice components of the GBM-LCCM and GP-LCCM are, similarly to the traditional LCCM, based on the same random utility theory. To test this hypothesis, we conducted a policy analysis study, using the case of the American University of Beirut. A base case scenario was developed and results showed that the three different models generate similar meaningful forecasts. Results also showed that around 41% of individuals would have a multimodal travel behavior in a given week under the base case scenario. However, only 14% and 25% of the weekly trips are expected to be made by shared-taxi and shuttle. It is acknowledged that these forecasts, which are based on a demand model derived from a stated preferences survey, may not fully materialize as respondents to SP surveys express their intention or preference which may not translate into actual market behavior. However, the models could not be calibrated to real market conditions since there is no information on the real weekly travel frequencies by different modes. As such, the forecasting results are indicative but not necessarily completely representative of the true ridership expected on the new modes.

Moreover, a sensitivity analysis of the demand for the potential services, with respect to their one-way fares, was conducted to investigate further the forecasting

performance of the different models. All models resulted in similar forecasts, to some extent. However, the main difference was depicted in the forecasts of the three-class GP-LCCM and the two-class LCCM. Overall, the GP-LCCM predicted higher (lower) number of weekly trips than the LCCM for low (high) fares. This difference in the forecasts might be attributed to the higher number of classes and/or the nonparametric nature of the class membership component of the GP-LCCM. Results also showed that the proposed models provide greater insights to the underlying behavioral heterogeneity within a population by identifying a larger number of latent classes that respond differently to new policies.

This policy analysis application has three main limitations. First, only one base case scenario was considered and a sensitivity analysis with respect to the one-way fares of the shared-taxi and shuttle was performed. Future work could explore whether the findings of this policy analysis application generalize to different applications or scenarios and different types of datasets. Second, since the LCCM and the two proposed models, GBM-LCCM and GP-LCCM, are characterized by similar utility-based class-specific choice models but different class membership models (i.e., different latent class models), future work could explore the impact of changes in the socio-economic variables that shape the latent classes on the forecasting performance of the different models. Finally, the GP-LCCM, which was able to estimate a higher number of classes than the LCCM, showed that the three identified classes have different modality styles. Consequently, different policies that take into consideration the different characteristics and behaviors of each class could be tested, rather than applying one general policy to the entire population.

# CHAPTER 8

## CONCLUSION

This chapter concludes the dissertation. Section 8.1 summarizes the dissertation topic, contributions, and main findings. Section 8.2 discusses limitations and directions for future research. Finally, Section 8.3 concludes.

### **8.1. Summary**

This dissertation investigated the feasibility of combining the strengths of machine learning methods and discrete choice models within hybrid econometric models. These strengths consist of the predictive power of unsupervised machine learning algorithms and their flexibility in detecting unobserved patterns, and the explanatory power and behavioral realism of discrete choice models. More specifically, the main objective of this dissertation was to integrate unsupervised machine learning algorithms into the Latent Class Choice Model structure to improve the overall model flexibility and discrete representation of heterogeneity without lessening the economic and behavioral interpretability of the choice models. We contributed to the literature by developing two hybrid models that satisfy the previously mentioned objective.

The first proposed model is a Gaussian-Bernoulli Mixture – Latent Class Choice Model (GBM-LCCM) that, similarly to traditional LCCM, consists of two components, a class membership model and a class-specific choice model. The class membership component, which predicts the probability of a decision-maker belonging to a specific latent class/cluster, is formulated as a parametric model-based mixture model with Gaussian and Bernoulli distributions instead of a random utility formulation.

Conditioned on the class assignments, the class-specific choice component formulates the probability of a particular alternative being chosen by a decision-maker using random utility models. An Expectation-Maximization (EM) algorithm was derived to estimate the proposed model. The model was evaluated using two different revealed and stated preferences datasets and was also compared to its traditional LCCM counterpart. Results showed that the GBM-LCCM has the capability to capture more complex discrete representation of heterogeneity than the LCCM (i.e., higher number of latent classes) and to improve the in-sample goodness of fit as well as the out-of-sample prediction accuracy without any economic or behavioral interpretability losses. Marginal effects and economic indicators can be easily inferred from the model and the latent classes (clusters) can be easily interpreted.

The second model is a Gaussian Process – Latent Class Choice Model (GP-LCCM) that also consists of two components, a class membership model and a class-specific choice model. The former is constructed as a Gaussian Process to model unobserved heterogeneity as discrete constructs (latent classes) while the latter estimates, similarly to the GBM-LCCM and traditional LCCM, the corresponding choice probabilities using random utility models. An EM algorithm was also derived and implemented to estimate/infer the parameters of the choice models and the hyper-parameters of the GP kernel function. The iterative nature of the EM algorithm enabled the use of the Laplace approximation method to infer the GP posterior for clustering purposes. The model was evaluated using three different datasets (two stated preferences and on revealed preferences) and benchmarked against the GBM-LCCM as well as the traditional LCCM. The proposed GP-LCCM outperformed the two benchmark models, LCCM and GBM-LCCM, in terms of in-sample goodness of fit and



out-of-sample generalization performance without compromising the economic and behavioral interpretability of the class-specific choice models. The model also demonstrated a higher degree of flexibility by capturing more complex discrete representation of heterogeneity.

A policy analysis was also conducted to explore the forecasting capabilities of the proposed models. Results showed that the two proposed models are capable of providing meaningful forecasts that are similar, to some extent, to the forecasts of the traditional LCCM. Results also showed similar order of change between the results of the proposed models and LCCM in terms of in-sample LL, out-of-sample LL, and forecasts. Finally, the results of a class-specific sensitivity analysis of the demand for different travel modes underlined the importance of the proposed models in identifying a higher number of classes than the LCCM, something that provides a more in-depth understanding of the different modality styles and behavioral heterogeneity within a population.

## **8.2. Limitations and Future Directions**

Both models displayed better flexibility and generalization performance than the LCCM. However, each has its advantages and drawbacks. The parametric nature of the model-based Gaussian-Bernoulli Mixture component of the GBM-LCCM guarantees a transparent interpretation of the latent classes, similarly to the random utility specification of the LCCM. However, the model assumes that the continuous and binary variables entering the class membership component are uncorrelated. On the contrary, the nonparametric nature of the Gaussian Process component of the GP-LCCM surpasses the correlation limitation of the GBM-LCCM, lessens the restrictive

parametric assumptions of both the LCCM and GBM-LCCM, and ensures higher flexibility and prediction accuracy. Nevertheless, the GP-LCCM has some drawbacks. First, the nonparametric nature of GP might lead to overfitting issues. This might be addressed by conducting a thorough investigation for a more suitable kernel or combination of kernels function. Second, finding an appropriate kernel(s) function imposes an additional burden on the modeler. Future work could explore ways to automate this task. Third, the nonparametric nature of GP affects the overall transparency of the latent classes. However, the interpretation of the classes can be locally achieved by means of model-agnostic techniques such as LIME. Finally, the Gaussian Process increases the computational cost of the model significantly. This limitation could be addressed by relying on Sparse Gaussian Processes (Titsias, 2009) and modern computational hardware (e.g., GPU).

In addition to model-specific limitations, several extensions can be considered to enhance the two models. First, although three different types of datasets with different sample sizes were used to evaluate the two proposed models, it would be worthwhile to investigate whether the findings of this dissertation generalize to different applications, specifications, and attribute transformations. Second, the two models account for discrete representations of heterogeneity. A natural extension is to incorporate random parameter distributions or mixture of distributions into the class-specific choice models to account for another layer of within-class random heterogeneity. Moreover, the two models can be extended to account for different aspects of heterogeneity such as systematic and random taste variations and variations in choice sets considered by decision-makers in addition to market segmentation (i.e., latent classes). This can be achieved by integrating within the two proposed models more machine learning

algorithms (e.g., artificial neural networks and decision trees), Bayesian nonparametrics (e.g., Dirichlet process mixtures models), and/or traditional statistics (e.g., mixture of distributions). Third, feedback from the class-specific choice models could be incorporated in the class membership model (e.g., through consumer surplus also known as logsum term) to account for preference endogeneity and the fact that preferences might be sensitive to changes in some alternative attributes (Vij & Walker, 2014). Fourth, future work can benchmark the two proposed models against a variety of advanced discrete choice models or other hybrid models that combine machine learning with behavioral theory. Fifth, ensemble learning could be applied by combining the predictions of the two proposed models in an ensemble with the aim of improving the prediction performance further. Sixth, the two models were able to identify a larger number of latent classes than the LCCM in the three applications considered in this dissertation. To better exploit such advantage, class-specific policy analysis might be tested instead of general policies as in Chapter 7. This would allow to develop different policies that account for the different characteristics, preferences, and behavior of each market segment. It could also lead to a higher degree of success in case the policies are implemented. Finally, future work can explore the benefits of using Bayesian Variational Inference (VI) estimation techniques as opposed to the Expectation-Maximization approach that was adopted in this dissertation. Discrete choice models are usually estimated using maximum likelihood strategies such as maximum simulated likelihood estimation and Expectation-Maximization techniques. However, relying on a fully-Bayesian approach as an alternative estimation strategy can guarantee several advantages such as the ability of generating the entire posterior distributions of all model parameters, performing automatic model specification (Rodrigues et al., 2020),

and dealing with missing data. However, Bayesian inference relies on approximate inference methods with Markov-chain Monte Carlo (MCMC) being the most used approximation method in the econometric literature (Danaf et al., 2019). Although MCMC methods are powerful sampling approaches for approximate posterior inference, they still suffer from high computational costs and several difficulties in assessing convergence especially when dealing with large and complex datasets (Bansal et al., 2020). More recently, variational inference methods that are arising in machine learning and Bayesian inference have been applied to a few mixed logit models (Bansal et al., 2020; Rodrigues, 2020). Variational inference methods transform the Bayesian inference approximation problem to an optimization problem and in doing so outperform the shortcomings of MCMC by reducing the computational time significantly without jeopardizing the model accuracy. Nevertheless, VI methods are still limited to mixed logit models with simple normal mixing distributions and still encounter several limitations such as the ability to approximate high complex posterior distributions and the scalability to large and complex datasets (Bansal et al., 2020). Future work can focus on estimating the GBM-LCCM and GP-LCCM through Bayesian VI techniques and on providing solutions to overcome the corresponding limitations of such methods.

### **8.3. Conclusion**

This dissertation has contributed to the efforts aiming at bridging the gap between machine learning and discrete choice models all while retaining the basic aspects of McFadden's original work to ensure a transparent behavioral and economic interpretation of the developed models. Finally, though the focus of the three

applications in this dissertation was on travel mode choice, the models can be applied to different scientific areas related to behavioral science such as economics, marketing research, and psychology, in addition to any application with a finite discrete choice set where it is believed that taste heterogeneity exists among decision-makers. It is hoped that these models and the abovementioned extensions (Section 8.2) could provide a stronger evidence base for the potential merits of the proposed models to the choice modeling and transportation planning communities.

## APPENDIX A: LONDON DATASET

This appendix provides all estimation results of the LCCM and GBM-LCCM models related to the London dataset presented in Chapters 4. Section A.1 presents the results of the first trial, Section A.2 presents the results of the second trial, and Section A.3 presents the results of the third trial. The specifications of the different trials are described in Section 4.3.1 of Chapter 4.

### **A.1. London Dataset – First Specification**

This part of the appendix presents the results of the first specification of the London dataset. Section A.1.1 presents the LCCM results and Section A.1.2 presents the GBM-LCCM results.

#### ***A.1.1. Latent Class Choice Models***

Tables A.1 and A.2 show the parameter estimates of the LCCM with two and three classes, respectively.

Table A.1: Estimation results of the first specification of the LCCM with two classes based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>
	<b>Class-Specific Choice Model</b>	
ASC (Car)	-0.904 (0.00)	2.34 (0.00)
Travel Time (PT) (minutes)	-0.0757 (0.00)	-0.0888 (0.00)
Travel Time (Car) (minutes)	-0.250 (0.00)	-0.127 (0.00)
Cost (PT) (£ gbp)	-0.205 (0.01)	0.0618 (0.14)
Cost (Car) (£ gbp)	-0.182 (0.08)	-0.209 (0.00)
<b>Parameter</b>	<b>Class Membership Model</b>	
ASC		-5.40 (0.00)
age		0.174 (0.00)
female		-0.287 (0.01)
license		2.04 (0.00)
car_own <sub>1</sub>		3.34 (0.00)
car_own <sub>2</sub>		4.72 (0.00)

Values within parentheses are p-values

Table A.2: Estimation results of the first specification of the LCCM with three classes based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>	<b>Class 3</b>
	<b>Class-Specific Choice Model</b>		
ASC (Car)	-4.72 (0.00)	3.99 (0.00)	1.88 (0.00)
Travel Time (PT) (minutes)	-0.0530 (0.00)	-0.171 (0.00)	-0.134 (0.00)
Travel Time (Car) (minutes)	-0.0493 (0.00)	-0.177 (0.00)	-0.249 (0.00)
Cost (PT) (£ gbp)	-0.0970 (0.25)	0.0748 (0.57)	-0.0413 (0.51)
Cost (Car) (£ gbp)	-0.0421 (0.32)	-0.260 (0.00)	-0.373 (0.00)
<b>Parameter</b>	<b>Class Membership Model</b>		
ASC		-11.85 (0.00)	-3.60 (0.00)
age		0.241 (0.00)	0.0358 (0.52)
female		-0.738 (0.01)	0.166 (0.26)
license		3.01 (0.00)	1.64 (0.00)
car_own <sub>1</sub>		8.06 (0.00)	2.57 (0.00)
car_own <sub>2</sub>		10.23 (0.00)	4.03 (0.00)

Values within parentheses are p-values

### *A.1.2. Gaussian-Bernoulli Mixture - Latent Class Choice Models*

This section presents the estimation results of the GBM-LCCMs with two, three, and four classes for both full and tied covariance structures. First, Section A.1.2.1 presents the models with a full covariance structure, then Section A.1.2.2 presents the models with a tied covariance structure.

#### A.1.2.1. Full Covariance

This section presents the estimation results of the full-GBM-LCCMs with two, three, and four classes.

Table A.3: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with two classes and a full covariance based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>
	<b>Class-Specific Choice Model</b>	
ASC (Car)	-0.00350 (0.00)	1.83 (0.00)
Travel Time (PT) (minutes)	-0.0879 (0.00)	-0.0831 (0.00)
Travel Time (Car)	-0.381 (0.00)	-0.121 (0.00)
Cost (PT) (£ gbp)	-0.428 (0.00)	-0.00270 (0.94)
Cost (Car) (£ gbp)	-0.211 (0.15)	-0.196 (0.00)

Values within parentheses are p-values



Table A.4: Class membership mean estimates of the first specification of the GBM-LCCM with two classes and a full covariance matrix based on the London dataset

Parameter		Class 1	Class 2
age*	Continuous	-0.261	0.264
female	Yes	0.523	0.413
	No*	0.477	0.587
license	Yes	0.547	0.959
	No*	0.453	0.041
car_own <sub>0</sub>	0*	0.436	0.076
car_own <sub>1</sub>	] 0 – 1 [	0.516	0.495
car_own <sub>2</sub>	≥ 1	0.048	0.429

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
Class 1, age:	0.880
Class 2, age:	0.983

Mixing coefficients:

- $\pi_1 = 0.503$
- $\pi_2 = 0.497$

Table A.5: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with three classes and a full covariance based on the London dataset

Parameter	Class 1	Class 2		Class 3
		Class-Specific Choice Model		
ASC (Car)	-0.774 (0.00)	1.59 (0.00)	1.87 (0.00)	
Travel Time (PT) (minutes)	-0.0783 (0.00)	-0.0661 (0.00)	-0.0875 (0.00)	
Travel Time (Car) (minutes)	-0.283 (0.00)	-0.110 (0.00)	-0.132 (0.00)	
Cost (PT) (£ gbp)	-0.272 (0.00)	-0.0288 (0.50)	-0.0323 (0.50)	
Cost (Car) (£ gbp)	-0.208 (0.04)	-0.154 (0.00)	-0.210 (0.00)	

Values within parentheses are p-values

Table A.6: Class membership mean estimates of the first specification of the GBM-LCCM with three classes and a full covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 3
<b>age*</b>	Continuous	-0.288	0.342	0.182
<b>female</b>	Yes	0.528	0.465	0.382
	No*	0.472	0.535	0.618
<b>license</b>	Yes	0.514	0.992	0.933
	No*	0.486	0.008	0.067
<b>car_own<sub>0</sub></b>	0*	0.525	0	0.044
<b>car_own<sub>1</sub></b>	] 0 – 1 [	0.451	0	0.956
<b>car_own<sub>2</sub></b>	≥ 1	0.025	1	0

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance      age  
Class 1, age:      0.866  
Class 2, age:      0.961  
Class 3, age:      0.986

Mixing coefficients:

- $\pi_1 = 0.464$
- $\pi_2 = 0.226$
- $\pi_3 = 0.310$

Table A.7: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with four classes and a full covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3	Class 4
ASC (Car)	-0.905 (0.00)	2.33 (0.00)	1.49 (0.00)	2.52 (0.00)
Travel Time (PT) (minutes)	-0.0730 (0.00)	-0.193 (0.00)	-0.0647 (0.00)	-0.105 (0.00)
Travel Time (Car) (minutes)	-0.268 (0.00)	-0.340 (0.00)	-0.106(0.00)	-0.110 (0.00)
Cost (PT) (£ gbp)	-0.257 (0.01)	-0.157 (0.12)	-0.0241 (0.57)	-0.121 (0.41)
Cost (Car) (£ gbp)	-0.192 (0.05)	-0.548 (0.00)	-0.154 (0.00)	-0.209 (0.00)

Values within parentheses are p-values

Table A.8: Class membership mean estimates of the first specification of the GBM-LCCM with four classes and a full covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 3	Class 4
age*	Continuous	-0.291	0.070	0.330	0.379
female	Yes	0.534	0.419	0.463	0.296
	No*	0.466	0.581	0.537	0.704
license	Yes	0.487	0.936	0.989	0.920
	No*	0.513	0.064	0.011	0.080
car_own <sub>0</sub>	0*	0.551	0.047	0.020	0.017
car_own <sub>1</sub>	] 0 – 1 [	0.425	0.953	0	0.983
car_own <sub>2</sub>	≥ 1	0.024	0	0.980	0

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
Class 1, age:	0.883
Class 2, age:	0.878
Class 3, age:	0.967
Class 4, age:	1.157

Mixing coefficients:

- $\pi_1 = 0.435$
- $\pi_2 = 0.248$
- $\pi_3 = 0.232$
- $\pi_4 = 0.086$

### A.1.2.2. Tied Covariance

This section presents the estimation results of the tied-GBM-LCCMs with two, three, and four classes.

Table A.9: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with two classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2
	Class-Specific Choice Model	
ASC (Car)	1.84 (0.00)	0.00460 (0.98)
Travel Time (PT) (minutes)	-0.0840 (0.00)	-0.0872 (0.00)
Travel Time (Car) (minutes)	-0.122 (0.00)	-0.375 (0.00)
Cost (PT) (£ gbp)	-0.00870 (0.81)	-0.407 (0.00)
Cost (Car) (£ gbp)	-0.197 (0.00)	-0.210 (0.14)

Values within parentheses are p-values

Table A.10: Class membership mean estimates of the first specification of the GBM-LCCM with two classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2
age*	Continuous	0.261	-0.257
female	Yes	0.413	0.524
	No*	0.587	0.476
license	Yes	0.960	0.546
	No*	0.040	0.454
car_own0	0*	0.076	0.435
car_own1	] 0 – 1 [	0.495	0.516
car_own2	≥ 1	0.429	0.049

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance      age  
age:                0.933

Mixing coefficients:

- $\pi_1 = 0.496$
- $\pi_2 = 0.504$

Table 11: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with three classes and a tied covariance based on the London dataset

Parameter	Class 1	Class-Specific Choice Model	
		Class 2	Class 3
ASC (Car)	1.88 (0.00)	1.60 (0.00)	-0.735 (0.00)
Travel Time (PT) (minutes)	-0.0886 (0.00)	-0.0663 (0.00)	-0.0784 (0.00)
Travel Time (Car) (minutes)	-0.132 (0.00)	-0.110 (0.00)	-0.285 (0.00)
Cost (PT) (£ gbp)	-0.0144 (0.77)	-0.0260 (0.55)	-0.263 (0.00)
Cost (Car) (£ gbp)	-0.210 (0.00)	-0.154 (0.00)	-0.207 (0.04)

Values within parentheses are p-values

Table A.12: Class membership mean estimates of the first specification of the GBM-LCCM with three classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 3
age*	Continuous	0.176	0.342	-0.284
female	Yes	0.382	0.465	0.528
	No*	0.618	0.535	0.472
license	Yes	0.934	0.992	0.513
	No*	0.066	0.008	0.487
car_own0	0*	0.044	0.000	0.524
car_own1	] 0 – 1 [	0.956	0.000	0.451
car_own2	≥ 1	0.000	1.000	0.025

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance      age  
age:              0.927

Mixing coefficients:

- $\pi_1 = 0.310$
- $\pi_2 = 0.226$
- $\pi_3 = 0.464$

Table A.13: Class-specific choice parameter estimates of the first specification of the GBM-LCCM with four classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3	Class 4
ASC (Car)	2.35 (0.00)	-0.858 (0.00)	2.51 (0.00)	1.52 (0.00)
Travel Time (PT) (minutes)	-0.178 (0.00)	-0.0751 (0.00)	-0.112 (0.00)	-0.0646 (0.00)
Travel Time (Car) (minutes)	-0.316 (0.00)	-0.284 (0.00)	-0.115 (0.00)	-0.106 (0.00)
Cost (PT) (£ gbp)	-0.102 (0.28)	-0.267 (0.01)	-0.106 (0.49)	-0.0206 (0.63)
Cost (Car) (£ gbp)	-0.492 (0.00)	-0.181 (0.06)	-0.207 (0.00)	-0.153 (0.00)

Values within parentheses are p-values

Table A.14: Class membership mean estimates of the first specification of the GBM-LCCM with four classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 3	Class 4
age*	Continuous	0.034	-0.285	0.447	0.332
female	Yes	0.426	0.533	0.279	0.463
	No*	0.574	0.467	0.721	0.537
license	Yes	0.919	0.488	0.945	0.990
	No*	0.081	0.512	0.055	0.010
car_own0	0*	0.056	0.551	0.014	0.018
car_own1	] 0 – 1 [	0.944	0.424	0.986	0
car_own2	≥ 1	0	0.025	0	0.982

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance      age  
age:              0.923

Mixing coefficients:

- $\pi_1 = 0.254$
- $\pi_2 = 0.431$
- $\pi_3 = 0.084$
- $\pi_4 = 0.231$

## A.2. London Dataset – Second Specification

This part of the appendix presents the results of the second specification of the London dataset. Section A.2.1 presents the LCCM results and Section A.2.2 presents the GBM-LCCM results.

### A.2.1. Latent Class Choice Models

Tables A.15 and A.16 show the parameter estimates of the LCCM with two and three classes, respectively.

Table A.15: Estimation results of the second specification of the LCCM with two classes based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>
	<b>Class-Specific Choice Model</b>	
ASC (Car)	-1.23 (0.00)	2.51 (0.00)
Travel Time – Access (PT)	-0.142 (0.00)	-0.0773 (0.00)
Travel Time – Rail/Bus (PT)	-0.0624 (0.00)	-0.0804 (0.00)
Travel Time – Interchange (PT)	-0.144 (0.00)	-0.109 (0.00)
Travel Time (Car)	-0.267 (0.00)	-0.128 (0.00)
Log Cost (PT)	-0.0542 (0.03)	0.0134 (0.55)
Cost (Car)	-0.246 (0.18)	-0.208 (0.00)
<b>Parameter</b>	<b>Class Membership Model</b>	
ASC		-5.30 (0.00)
age		0.178 (0.00)
female		-0.298 (0.01)
license		2.02 (0.00)
car_own <sub>1</sub>		3.24 (0.00)
car_own <sub>2</sub>		4.59 (0.00)

Values within parentheses are p-values

Travel Time variables are in minutes

Cost variables are in Pound Sterling (£ gbp)

Table A.16: Estimation results of the second specification of the LCCM with three classes based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>	<b>Class 3</b>
<b>Class-Specific Choice Model</b>			
ASC (Car)	-4.12 (0.00)	5.24 (0.00)	1.38 (0.00)
Travel Time – Access (PT)	-0.0831 (0.00)	-0.0618 (0.19)	-0.130 (0.00)
Travel Time – Rail/Bus (PT)	-0.0380 (0.00)	-0.174 (0.00)	-0.121 (0.00)
Travel Time – Interchange (PT)	-0.144 (0.00)	-0.176 (0.00)	-0.150 (0.00)
Travel Time (Car)	-0.0360 (0.00)	-0.203 (0.00)	-0.240 (0.00)
Log Cost (PT)	0.188 (0.00)	-0.256 (0.03)	-0.101 (0.03)
Cost (Car)	-0.00970 (0.79)	-0.300 (0.00)	-0.363 (0.00)
<b>Class Membership Model</b>			
ASC		-16.24 (0.00)	-3.56 (0.00)
age		0.338 (0.00)	0.0426 (0.43)
female		-0.800 (0.01)	0.165 (0.25)
license		3.70 (0.00)	1.58 (0.00)
car_own <sub>1</sub>		11.28 (0.01)	2.58 (0.00)
car_own <sub>2</sub>		13.48 (0.00)	4.07 (0.00)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)



### *A.2.2. Gaussian-Bernoulli Mixture - Latent Class Choice Models*

This section presents the estimation results of the GBM-LCCMs with two, three, and four classes for both full and tied covariance structures. First, Section A.2.2.1 presents the models with a full covariance structure, and then Section A.2.2.2 presents the models with a tied covariance structure.

#### A.2.2.1. Full Covariance

This section presents the estimation results of the full-GBM-LCCMs with two, three, and four classes.

Table A.17: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with two classes and a full covariance based on the London dataset

<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>
	<b>Class-Specific Choice Model</b>	
ASC (Car)	1.76 (0.00)	-1.02 (0.00)
Travel Time – Access (PT)	-0.0801 (0.00)	-0.172 (0.00)
Travel Time – Rail/Bus (PT)	-0.0770 (0.00)	-0.0609 (0.00)
Travel Time – Interchange (PT)	-0.102 (0.00)	-0.172 (0.00)
Travel Time (Car)	-0.121 (0.00)	-0.374 (0.00)
Log Cost (PT)	-0.0229 (0.20)	-0.160 (0.00)
Cost (Car)	-0.197 (0.00)	-0.219 (0.17)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)

Table A.18: Class membership mean estimates of the second specification of the GBM-LCCM with two classes and a full covariance matrix based on the London dataset

	<b>Parameter</b>	<b>Class 1</b>	<b>Class 2</b>
age*	Continuous	0.266	-0.261
female	Yes	0.413	0.524
	No*	0.587	0.477
license	Yes	0.958	0.550
	No*	0.042	0.450
car_own0	0*	0.076	0.434
car_own1	] 0 – 1 [	0.494	0.517
car_own2	≥ 1	0.429	0.049

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
Class 1, age:	0.991
Class 2, age:	0.872

Mixing coefficients:

- $\pi_1 = 0.495$
- $\pi_2 = 0.505$

Table A.19: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with three classes and a full covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3
	Class-Specific Choice Model		
ASC (Car)	-1.76 (0.00)	1.27 (0.00)	2.01 (0.00)
Travel Time – Access (PT)	-0.145 (0.00)	-0.0736 (0.00)	-0.0667 (0.00)
Travel Time – Rail/Bus (PT)	-0.0645 (0.00)	-0.0660 (0.00)	-0.0730 (0.00)
Travel Time – Interchange (PT)	-0.124 (0.00)	-0.0656 (0.00)	-0.146 (0.00)
Travel Time (Car)	-0.286 (0.00)	-0.110 (0.00)	-0.130 (0.00)
Log Cost (PT)	-0.139 (0.00)	-0.0463 (0.03)	-0.0276 (0.26)
Cost (Car)	-0.208 (0.04)	-0.156 (0.00)	-0.211 (0.00)

Values within parentheses are p-values  
 Travel Time variables are in minutes  
 Cost variables are in Pound Sterling (£ gbp)

Table A.20: Class membership mean estimates of the second specification of the GBM-LCCM with three classes and a full covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 2
age*	Continuous	-0.291	0.343	0.183
female	Yes	0.529	0.465	0.382
	No*	0.471	0.535	0.618
license	Yes	0.514	0.992	0.931
	No*	0.486	0.008	0.069
car_own0	0*	0.525	0.000	0.045
car_own1	] 0 – 1 [	0.450	0.000	0.955
car_own2	≥ 1	0.025	1.000	0.000

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance	age
Class 1, age:	0.860
Class 2, age:	0.962
Class 3, age:	0.990

Mixing coefficients:

- $\pi_1 = 0.463$
- $\pi_2 = 0.226$
- $\pi_3 = 0.311$

Table A.21: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with four classes and a full covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3	Class 4
ASC (Car)	3.41 (0.00)	1.18 (0.00)	-1.81 (0.00)	1.00 (0.02)
Travel Time – Access (PT)	-0.0430 (0.26)	-0.0769 (0.00)	-0.123 (0.00)	-0.204 (0.00)
Travel Time – Rail/Bus (PT)	-0.0526 (0.00)	-0.0654 (0.00)	-0.0533 (0.00)	-0.315 (0.00)
Travel Time – Interchange (PT)	-0.214 (0.00)	-0.0639 (0.00)	-0.112 (0.00)	-0.344 (0.00)
Travel Time (Car)	-0.0761 (0.00)	-0.106 (0.00)	-0.208 (0.00)	-0.488 (0.00)
Log Cost (PT)	0.0779 (0.33)	-0.0368 (0.07)	-0.0832 (0.00)	-0.430 (0.00)
Cost (Car)	-0.210 (0.00)	-0.159 (0.00)	-0.227 (0.02)	-1.15 (0.06)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)

Table A.22: Class membership mean estimates of the second specification of the GBM-LCCM with four classes and a full covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 3	Class 4
age*	Continuous	0.274	0.326	-0.300	0.151
female	Yes	0.351	0.463	0.535	0.388
	No*	0.649	0.537	0.465	0.612
license	Yes	0.858	0.987	0.492	0.993
	No*	0.142	0.013	0.508	0.007
car_own0	0*	0.028	0.028	0.543	0.010
car_own1	] 0 – 1 [	0.972	0	0.436	0.990
car_own2	≥ 1	0	0.972	0.022	0

\*: base category

\*\*: continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance  
Class 1, age: 1.265  
Class 2, age: 0.967  
Class 3, age: 0.862  
Class 4, age: 0.846

Mixing coefficients:  
•  $\pi_1 = 0.097$   
•  $\pi_2 = 0.234$   
•  $\pi_3 = 0.453$   
•  $\pi_4 = 0.216$

### A.2.2.2. Tied Covariance

This section presents the estimation results of the full-GBM-LCCMs with two, three, and four classes.

Table A.23: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with two classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2
	Class-Specific Choice Model	
ASC (Car)	1.84 (0.00)	-0.902 (0.00)
Travel Time – Access (PT)	-0.0813 (0.00)	-0.169 (0.00)
Travel Time – Rail/Bus (PT)	-0.0779 (0.00)	-0.0600 (0.00)
Travel Time – Interchange (PT)	-0.103 (0.00)	-0.172 (0.00)
Travel Time (Car)	-0.122 (0.00)	-0.368 (0.00)
Log Cost (PT)	-0.0115 (0.53)	-0.142 (0.00)
Cost (Car)	-0.198 (0.00)	-0.217 (0.16)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)

Table A.24: Class membership mean estimates of the second specification of the GBM-LCCM with two classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2
age*	Continuous	0.261	-0.256
female	Yes	0.412	0.524
	No*	0.588	0.476
license	Yes	0.959	0.549
	No*	0.041	0.451
car_own0	0*	0.076	0.434
car_own1	] 0 – 1 [	0.495	0.516
car_own2	≥ 1	0.429	0.050

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

#### Covariance

age: 0.933

#### Mixing coefficients:

- $\pi_1 = 0.494$
- $\pi_2 = 0.506$

Table A.25: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with three classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3
	Class-Specific Choice Model		
ASC (Car)	-1.64 (0.00)	2.11 (0.00)	1.29 (0.00)
Travel Time – Access (PT)	-0.144 (0.00)	-0.0686 (0.00)	-0.0736 (0.00)
Travel Time – Rail/Bus (PT)	-0.0638 (0.00)	-0.0738 (0.00)	-0.0663 (0.00)
Travel Time – Interchange (PT)	-0.126 (0.00)	-0.147 (0.00)	-0.0653 (0.00)
Travel Time (Car)	-0.287 (0.00)	-0.131 (0.00)	-0.110 (0.00)
Log Cost (PT)	-0.128 (0.00)	-0.0114 (0.65)	-0.0449 (0.04)
Cost (Car)	-0.207 (0.04)	-0.211 (0.00)	-0.156 (0.00)

Values within parentheses are p-values  
 Travel Time variables are in minutes  
 Cost variables are in Pound Sterling (£ gbp)

Table A.26: Class membership mean estimates of the second specification of the GBM-LCCM with three classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2	Class 2
age*	Continuous	-0.285	0.176	0.343
female	Yes	0.529	0.381	0.465
	No*	0.471	0.619	0.535
license	Yes	0.513	0.933	0.992
	No*	0.487	0.067	0.008
car_own <sub>0</sub>	0*	0.525	0.045	0.000
car_own <sub>1</sub>	] 0 – 1 [	0.450	0.955	0.000
car_own <sub>2</sub>	≥ 1	0.025	0.000	1.000

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
age:	0.926

Mixing coefficients:

- $\pi_1 = 0.463$
- $\pi_2 = 0.312$
- $\pi_3 = 0.226$

Table A.27: Class-specific choice parameter estimates of the second specification of the GBM-LCCM with four classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2	Class 3	Class 4
	<b>Class-Specific Choice Model</b>			
ASC (Car)	2.04 (0.00)	-1.84 (0.00)	3.46 (0.00)	1.20 (0.00)
Travel Time – Access (PT)	-0.120 (0.00)	-0.141 (0.00)	-0.0560 (0.25)	-0.0764 (0.00)
Travel Time – Rail/Bus (PT)	-0.171 (0.00)	-0.0589 (0.00)	-0.0907 (0.00)	-0.0649 (0.00)
Travel Time – Interchange (PT)	-0.251 (0.00)	-0.114 (0.00)	-0.226 (0.00)	-0.0635 (0.00)
Travel Time (Car)	-0.329 (0.00)	-0.271 (0.00)	-0.106 (0.00)	-0.106 (0.00)
Log Cost (PT)	-0.199 (0.00)	-0.124 (0.00)	0.0234 (0.82)	-0.0356 (0.09)
Cost (Car)	-0.563 (0.01)	-0.189 (0.05)	-0.223 (0.00)	-0.156 (0.00)

Values within parentheses are p-values  
 Travel Time variables are in minutes  
 Cost variables are in Pound Sterling (£ gbp)

Table A.28: Class membership mean estimates of the second specification of the GBM-LCCM with four classes and a tied covariance matrix based on the London dataset

Parameter	Class 1	Class 2	Class 3	Class 4
age* Continuous	0.053	-0.294	0.415	0.331
female Yes	0.428	0.535	0.277	0.463
female No*	0.572	0.465	0.723	0.537
license Yes	0.922	0.487	0.939	0.989
license No*	0.078	0.514	0.061	0.011
car_own0 0*	0.051	0.554	0.015	0.021
car_own1 ] 0 – 1 [	0.949	0.421	0.985	0.000
car_own2 ≥ 1	0.000	0.025	0.000	0.979

\*: base category

\*\*: continuous variable that is standardized to have a mean of 0 and standard deviation of 1

Covariance  
 age: 0.926

Mixing coefficients:  
 •  $\pi_1 = 0.250$   
 •  $\pi_2 = 0.430$   
 •  $\pi_3 = 0.089$   
 •  $\pi_4 = 0.231$

### A.3. London Dataset – Third Specification

This part of the appendix presents the results of the third specification of the London dataset. Section A.3.1 presents the GBM-LCCM results.

#### A.3.1. *Gaussian-Bernoulli Mixture - Latent Class Choice Models*

This section presents the estimation results of the GBM-LCCMs with two classes for both full and tied covariance structures. First, Section A.2.2.1 presents the GBM-LCCM with a full covariance structure, and then Section A.2.2.2 presents the GBM-LCCM with a tied covariance structure.

##### A.3.1.1. Full Covariance

Table A.29: Class-specific choice parameter estimates of the third specification of the GBM-LCCM with two classes and a full covariance based on the London dataset

Parameter	Class 1	Class 2
	Class-Specific Choice Model	
ASC (Car)	0.144 (0.73)	4.54 (0.00)
AM Peak (Car)	0.838 (0.00)	0.923 (0.00)
Inter Peak (Car)	0.702 (0.02)	1.72 (0.00)
Peak (Car)	1.48 (0.00)	0.978 (0.00)
Week Days (Car)	-0.234 (0.54)	-2.01 (0.01)
Saturday (Car)	0.375 (0.41)	-0.957 (0.22)
Winter (Car)	-0.562 (0.02)	0.166 (0.25)
Variability (Car)	-5.51 (0.00)	-5.15 (0.00)
Travel Time (PT)	-0.0674 (0.00)	-0.0658 (0.00)
Travel Time (Car)	-0.277 (0.00)	-0.0886 (0.00)
Cost (PT)	-0.454 (0.00)	-0.0428 (0.23)
Cost (Car)	-0.129 (0.46)	-0.145 (0.00)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)



Table A.30: Class membership mean estimates of the third specification of the GBM-LCCM with two classes and a full covariance matrix based on the London dataset

<b>Parameter</b>		<b>Class 1</b>	<b>Class 2</b>
age*	Continuous	-0.270	0.247
female	Yes	0.534	0.409
	No*	0.467	0.591
license	Yes	0.526	0.958
	No*	0.474	0.042
car_own0	0*	0.439	0.091
car_own1	] 0 – 1 [	0.517	0.495
car_own2	≥ 1	0.044	0.415

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
Class 1, age:	0.881
Class 2, age:	0.247

Mixing coefficients:

- $\pi_1 = 0.478$
- $\pi_2 = 0.522$

### A.3.1.2. Tied Covariance

Table A.31: Class-specific choice parameter estimates of the third specification of the GBM-LCCM with two classes and a tied covariance based on the London dataset

Parameter	Class 1	Class 2
	<b>Class-Specific Choice Model</b>	
ASC (Car)	0.156 (0.71)	4.55(0.00)
AM Peak (Car)	0.829 (0.00)	0.931 (0.00)
Inter Peak (Car)	0.681 (0.02)	1.74 (0.00)
Peak (Car)	1.46 (0.00)	0.983 (0.00)
Week Days (Car)	-0.225 (0.55)	-2.00 (0.01)
Saturday (Car)	0.386 (0.39)	-0.954 (0.22)
Winter (Car)	-0.553 (0.02)	0.161 (0.26)
Variability (Car)	-5.46 (0.00)	-5.17 (0.00)
Travel Time (PT)	-0.0671 (0.00)	-0.0662 (0.00)
Travel Time (Car)	-0.274 (0.00)	-0.0887 (0.00)
Cost (PT)	-0.430 (0.00)	-0.0338 (0.35)
Cost (Car)	-0.125 (0.45)	-0.146 (0.00)

Values within parentheses are p-values  
Travel Time variables are in minutes  
Cost variables are in Pound Sterling (£ gbp)

Table A.32: Class membership mean estimates of the third specification of the GBM-LCCM with two classes and a tied covariance matrix based on the London dataset

Parameter		Class 1	Class 2
age*	Continuous	-0.266	0.243
female	Yes	0.534	0.409
	No*	0.466	0.591
license	Yes	0.526	0.959
	No*	0.474	0.041
car_own <sub>0</sub>	0*	0.439	0.091
car_own <sub>1</sub>	] 0 – 1 [	0.517	0.495
car_own <sub>2</sub>	≥ 1	0.044	0.414

\*: base category

\*\* : continuous variable that is standardized to have a mean of 0 and standard deviation of 1

<u>Covariance</u>	age
age:	0.935

Mixing coefficients:

- $\pi_1 = 0.478$
- $\pi_2 = 0.522$

## APPENDIX B: AUB DATASET

This appendix provides all estimation results of the GBM-LCCM models related to the AUB dataset presented in Chapter 4. Sections B.1, B.2, B.3, and B.4 present the results of the models with full, tied, diagonal, and spherical covariance structures, respectively.

### B.1. Full Covariance

Table B.1: GBM-LCCM with two classes and a full covariance based on the AUB dataset

Parameter	Class 1 Class-specific choice model	Class 2	Parameter	Class 1	Class 2
				Class membership model	
$C_{car1}$	-1.91 (0.02)	0.315 (0.01)	$\pi$	0.573	0.427
$C_{car2}$	-1.84 (0.00)	0.257 (0.02)	$\mu_{Age}$	0.231	-0.310
$C_{car3}$	-2.22 (0.00)	0.467 (0.00)	$\mu_{Grade}$	0.055	-0.074
$C_{car4}$	-2.32 (0.01)	-0.613 (0.00)	$\mu_{C/D}$	0.336	-0.452
$C_{ST1}$	-1.41 (0.12)	-0.512 (0.00)	$\mu_{Nb}$	0.067	-0.090
$C_{ST2}$	-1.80 (0.00)	-0.201 (0.17)			
$C_{ST3}$	-1.24 (0.01)	-0.111 (0.59)			
$C_{ST4}$	-3.54 (0.00)	-0.339 (0.26)			
$C_{ST5}$	-0.198 (0.43)	-0.109 (0.75)			
$C_{SH1}$	-2.94 (0.00)	-0.194 (0.11)			
$C_{SH2}$	-3.41 (0.00)	0.445 (0.00)			
$C_{SH3}$	-2.61 (0.00)	0.680 (0.00)			
$C_{SH4}$	-4.79 (0.00)	0.367 (0.08)			
$C_{SH5}$	-1.53 (0.00)	0.499 (0.0465)			
$Cost_{Car}$	-0.0427 (0.00)	-0.0504 (0.00)			
$Cost_{ST}$	-0.0988 (0.00)	-0.113 (0.00)			
$Cost_{SH}$	-0.0393 (0.00)	-0.107 (0.00)			
$TT_{Car}$	-0.411 (0.00)	-0.625 (0.00)			
$TT_{ST}$	-0.370 (0.00)	-0.615 (0.00)			
$TT_{SH}$	-0.258 (0.00)	-0.334 (0.00)			
<b>Headway</b>	<b>-0.0455 (0.63)</b>	<b>-0.508 (0.00)</b>			

Values within parentheses are p-values

Cost variables are in 1,000 L.L

Travel Time and Headway variables are in hours

Table B.2: Full covariance matrix of the GBM-LCCM with two classes based on the AUB dataset

Covariance	Age	Grade	C/D	Nb
Class 1, Age	0.114	0.902	0.040	-0.035
Class1, Grade	0.036	0.040	1.096	-0.180
Class 1, C/D	1.075	0.114	0.036	0.127
Class 1, Nb	0.127	-0.035	-0.180	1.116
Class 2, Age	0.451	0.965	0.022	-0.149
Class 2, Grade	0.033	0.022	0.862	-0.224
Class 2, C/D	0.543	0.451	0.033	-0.027
Class 2, Nb	-0.027	-0.149	-0.224	0.830

## B.2. Tied Covariance

Table B.3: GBM-LCCM with two classes and a tied covariance based on the AUB dataset

Parameter	Class 1 Class-specific choice model	Class 2	Parameter	Class 1	Class 2
				Class membership model	
$C_{car1}$	-2.50 (0.00)	0.361 (0.00)	$\pi$	0.575	0.425
$C_{car2}$	-2.04 (0.00)	0.290 (0.01)	$\mu_{Age}$	0.303	-0.409
$C_{car3}$	-2.39 (0.00)	0.508 (0.00)	$\mu_{Grade}$	0.225	-0.303
$C_{car4}$	-3.08 (0.00)	-0.430 (0.00)	$\mu_{C/D}$	0.0459	-0.062
$C_{ST1}$	-1.62 (0.04)	-0.465 (0.00)	$\mu_{Nb}$	0.0513	-0.0693
$C_{ST2}$	-2.09 (0.00)	-0.174 (0.24)			
$C_{ST3}$	-1.08 (0.03)	-0.108 (0.61)			
$C_{ST4}$	-3.15 (0.00)	-0.347 (0.25)			
$C_{ST5}$	-0.159 (0.53)	-0.209 (0.55)			
$C_{SH1}$	-2.30 (0.00)	-0.286 (0.02)			
$C_{SH2}$	-3.03 (0.00)	0.403 (0.00)			
$C_{SH3}$	-2.29 (0.00)	0.661 (0.00)			
$C_{SH4}$	-4.02 (0.00)	0.354 (0.11)			
$C_{SH5}$	-1.52 (0.00)	0.379 (0.15)			
$Cost_{Car}$	-0.0442(0.00)	-0.0462 (0.00)			
$Cost_{ST}$	-0.101 (0.00)	-0.110 (0.00)			
$Cost_{SH}$	-0.0401 (0.00)	-0.0993 (0.00)			
$TT_{Car}$	-0.409 (0.00)	-0.653 (0.00)			
$TT_{ST}$	-0.372 (0.00)	-0.641 (0.00)			
$TT_{SH}$	-0.252 (0.00)	-0.384 (0.00)			
Headway	-0.0442 (0.64)	-0.561 (0.00)			

Values within parentheses are p-values

Cost variables are in 1,000 L.L

Travel Time and Headway variables are in hours

Table B.4: Tied covariance matrix of the GBM-LCCM with two classes based on the AUB dataset

Covariance	Age	Grade	C/D	Nb
Age	0.270	0.932	0.036	-0.078
Grade	0.041	0.036	0.997	-0.197
C/D	0.876	0.270	0.041	0.070
Nb	0.070	-0.078	-0.197	0.996

### B.3. Diagonal Covariance

Table B.5: GBM-LCCM with two classes and a diagonal covariance based on the AUB dataset

Parameter	Class 1 Class-specific choice model	Class 2	Parameter	Class 1	Class 2
				Class membership model	
$C_{car1}$	-1.95 (0.02)	0.312 (0.01)	$\pi$	0.569	0.431
$C_{car2}$	-1.90 (0.00)	0.251 (0.02)	$\mu_{Age}$	0.330	-0.436
$C_{car3}$	-2.51 (0.00)	0.477 (0.00)	$\mu_{Grade}$	0.244	-0.321
$C_{car4}$	-2.74 (0.00)	-0.506 (0.00)	$\mu_{C/D}$	0.049	-0.065
$C_{ST1}$	-1.54 (0.09)	-0.473 (0.00)	$\mu_{Nb}$	0.064	-0.085
$C_{ST2}$	-1.97 (0.00)	-0.167 (0.252)			
$C_{ST3}$	-1.12 (0.02)	-0.106 (0.61)			
$C_{ST4}$	-3.66 (0.00)	-0.314 (0.299)			
$C_{ST5}$	-0.183(0.47)	-0.120 (0.72)			
$C_{SH1}$	-2.65 (0.00)	-0.260 (0.03)			
$C_{SH2}$	-3.27 (0.00)	0.405 (0.00)			
$C_{SH3}$	-2.54 (0.00)	0.653 (0.00)			
$C_{SH4}$	-4.69 (0.00)	0.342 (0.11)			
$C_{SH5}$	-1.53 (0.00)	0.460 (0.07)			
$Cost_{Car}$	-0.0431 (0.00)	-0.0479 (0.00)			
$Cost_{ST}$	-0.0993 (0.00)	-0.113 (0.00)			
$Cost_{SH}$	-0.0401 (0.00)	-0.0999 (0.00)			
$TT_{Car}$	-0.406 (0.00)	-0.641 (0.00)			
$TT_{ST}$	-0.368 (0.00)	-0.626 (0.00)			
$TT_{SH}$	-0.250 (0.00)	-0.369 (0.00)			
Headway	-0.0521 (0.59)	-0.542 (0.00)			

Values within parentheses are p-values

Cost variables are in 1,000 L.L

Travel Time and Headway variables are in hours

Table B.6: Diagonal covariance matrix of the GBM-LCCM with two classes based on the AUB dataset

	Covariance	Age	Grade	C/D	Nb
Class 1		1.067	0.901	1.087	1.122
Class 2		0.577	0.949	0.877	0.827

#### B.4. Spherical Covariance

Table B.7: GBM-LCCM with two classes and a spherical covariance based on the AUB dataset

Parameter	Class 1	Class 2	Parameter	Class 1	Class 2
	Class-specific choice model			Class membership model	
$C_{car1}$	-2.17 (0.00)	0.332 (0.00)	$\pi$	0.572	0.428
$C_{car2}$	-1.92 (0.00)	0.261 (0.02)	$\mu_{Age}$	0.318	-0.425
$C_{car3}$	-2.47 (0.00)	0.490 (0.00)	$\mu_{Grade}$	0.242	-0.323
$C_{car4}$	-2.94 (0.00)	-0.463 (0.00)	$\mu_{C/D}$	0.052	-0.069
$C_{ST1}$	-1.57 (0.06)	-0.466 (0.00)	$\mu_{Nb}$	0.058	-0.078
$C_{ST2}$	-2.02 (0.00)	-0.168 (0.25)			
$C_{ST3}$	-1.08 (0.02)	-0.117 (0.57)			
$C_{ST4}$	-3.32 (0.00)	-0.331 (0.275)			
$C_{ST5}$	-0.171 (0.50)	-0.180 (0.60)			
$C_{SH1}$	-2.51 (0.00)	-0.270 (0.02)			
$C_{SH2}$	-3.17 (0.00)	0.404 (0.00)			
$C_{SH3}$	-2.42 (0.00)	0.652 (0.00)			
$C_{SH4}$	-4.41 (0.00)	0.342 (0.11)			
$C_{SH5}$	-1.52 (0.00)	0.413 (0.11)			
$Cost_{Car}$	-0.0434 (0.00)	-0.0480 (0.00)			
$Cost_{ST}$	-0.100 (0.00)	-0.111 (0.00)			
$Cost_{SH}$	-0.0404 (0.00)	-0.0995 (0.00)			
$TT_{Car}$	-0.408 (0.00)	-0.644 (0.00)			
$TT_{ST}$	-0.369 (0.00)	-0.635 (0.00)			
$TT_{SH}$	-0.252 (0.00)	-0.374 (0.00)			
Headway	0.0486 (0.61)	-0.554 (0.00)			

Values within parentheses are p-values

Cost variables are in 1,000 L.L

Travel Time and Headway variables are in hours

#### Covariance

Class 1	1.042
Class 2	0.814

Table B.8: GBM-LCCM with three classes and a spherical covariance based on the AUB dataset

Parameter	Class 1	Class 2	Class 3
	Class-specific choice model		
$C_{car1}$	-2.24 (0.00)	0.444 (0.00)	-0.102 (0.78)
$C_{car2}$	-1.82 (0.00)	0.290 (0.02)	0.327 (0.25)
$C_{car3}$	-2.11 (0.00)	0.677 (0.00)	-0.0412 (0.88)
$C_{car4}$	-2.90 (0.00)	-0.224 (0.18)	-1.26 (0.00)
$C_{ST1}$	-1.61 (0.01)	-0.331 (0.01)	-1.11 (0.00)
$C_{ST2}$	-2.31 (0.00)	-0.00930 (0.95)	-0.907 (0.03)
$C_{ST3}$	-1.07 (0.02)	0.0262 (0.91)	-0.938 (0.09)
$C_{ST4}$	-3.23 (0.00)	-0.073 (0.82)	-1.78 (0.08)
$C_{ST5}$	-0.181 (0.47)	-0.284 (0.48)	-0.24 (0.76)
$C_{SH1}$	-2.19 (0.00)	-0.240 (0.07)	-0.557 (0.08)
$C_{SH2}$	-3.17 (0.00)	0.556 (0.00)	-0.535 (0.11)
$C_{SH3}$	-2.12 (0.00)	0.862 (0.00)	-0.959 (0.03)
$C_{SH4}$	-4.38 (0.00)	0.627 (0.01)	-1.63 (0.00)
$C_{SH5}$	-1.45 (0.00)	0.601 (0.04)	-1.15 (0.05)
$Cost_{car}$	-0.0451 (0.00)	-0.0707 (0.00)	-0.0172 (0.11)
$Cost_{ST}$	-0.0991 (0.00)	-0.107 (0.00)	-0.123 (0.00)
$Cost_{SH}$	-0.0421 (0.00)	-0.0832 (0.00)	-0.120 (0.00)
$TT_{car}$	-0.410 (0.00)	-0.717 (0.00)	-0.614 (0.00)
$TT_{ST}$	-0.384 (0.00)	-0.777 (0.00)	-0.349 (0.01)
$TT_{SH}$	-0.259 (0.00)	-0.519 (0.00)	-0.107 (0.26)
Headway	-0.0114 (0.91)	-0.757 (0.00)	-0.380 (0.06)
Parameter	Class membership model		
$\pi$	0.570	0.339	0.091
$\mu_{Grade}$	0.258	-0.172	-0.981
$\mu_{C/D}$	0.0456	-0.222	0.540
$\mu_{Age}$	0.329	-0.314	-0.897
$\mu_{Nb}$	0.0778	0.0861	-0.810

Values within parentheses are p-values

Cost variables are in 1,000 L.L

Travel Time and Headway variables are in hours

Covariance

Class 1	1.044
Class 2	0.889
Class 3	0



## APPENDIX C: SWISSMETRO DATASET

This appendix provides the estimation results of the best models related to the Swissmetro dataset presented in Chapter 6. Section C.1 presents the best LCCM and Section C.2 presents the best GP-LCCM.

### C.1. Latent Class Choice Model

This section shows the estimation results of the LCCM with five latent classes.

Table C.1: Class-specific choice parameter estimates of the LCCM with five classes based on the Swissmetro dataset

Parameter	Class 1	Class 2	Class 3	Class 4	Class 5
ASC (Train)	0.811 (0.00)	1.19 (0.00)	-1.64 (0.00)	-0.254 (0.00)	-2.35 (0.00)
ASC (Car)	4.38 (0.00)	-0.126 (0.41)	-4.52 (0.11)	1.72 (0.00)	-1.27 (0.00)
Travel Time	-2.61 (0.08)	-0.0734 (0.00)	-0.129 (0.00)	-4.09 (0.17)	-4.36 (0.00)
Travel Cost	-1.42 (0.00)	-1.04 (0.00)	-0.502 (0.00)	-3.43 (0.00)	-3.63 (0.00)

Values within parentheses are p-values

Table C.2: Class membership parameter estimates of the LCCM with five classes based on the Swissmetro dataset

Parameter	Class 2	Class 3	Class 4	Class 5
ASC	-3.02 (0.00)	-2.32 (0.06)	-3.73 (0.08)	-3.30 (0.00)
24 < AGE ≤ 39	1.22 (0.00)	2.13 (0.00)	2.98 (0.00)	3.05 (0.00)
39 < AGE ≤ 54	0.00680 (0.99)	1.17 (0.04)	2.09 (0.00)	2.20 (0.00)
54 < AGE ≤ 65	1.28 (0.01)	0.667 (0.37)	1.95 (0.00)	2.56 (0.00)
AGE > 65	2.15 (0.00)	0.658 (0.49)	1.54 (0.02)	1.98 (0.01)
50 ≤ INCOME ≤ 100	-1.13 (0.00)	-0.397 (0.38)	-0.325 (0.27)	-0.379 (0.20)
INCOME > 100	-1.37 (0.01)	-0.590 (0.28)	-0.258 (0.39)	-0.381 (0.21)
M_INCOME	0.00250 (0.99)	-0.345 (0.55)	-0.486 (0.31)	-0.406 (0.36)
MALE	-0.775 (0.01)	-0.580 (0.08)	0.261 (0.31)	0.179 (0.48)
FIRST	0.196 (0.53)	0.258 (0.47)	0.626 (0.00)	0.985 (0.00)
LUGGAGE : 0 piece	1.26 (0.21)	-1.01 (0.33)	2.80 (0.18)	1.47 (0.05)
LUGGAGE : 1 piece	1.55 (0.11)	-0.411 (0.67)	2.03 (0.33)	0.599 (0.43)
PURPUSE: Commuter	3.28 (0.00)	4.76 (0.00)	1.15 (0.00)	2.65 (0.00)
PURPUSE: Shopping	3.55 (0.00)	4.66 (0.00)	1.68 (0.00)	2.07 (0.00)
PURPUSE: Business	2.25 (0.00)	2.01 (0.00)	0.398 (0.00)	1.05 (0.00)

Values within parentheses are p-values

## C.2. Gaussian Process - Latent Class Choice Model

This section shows the estimation results of the GP-LCCM with seven latent classes.

Table C.3: Class-specific choice parameter estimates of the GP-LCCM with seven classes based on the Swissmetro dataset

Parameter	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
ASC (Train)	1.71 (0.00)	-0.575 (0.00)	0.281 (0.00)	0.625 (0.00)	-2.06 (0.00)	-3.38 (0.00)	2.15 (0.00)
ASC (Car)	-1.31 (0.00)	-3.59 (0.00)	2.27 (0.00)	5.29 (0.00)	-0.292 (0.00)	-2.39 (0.00)	1.85 (0.00)
Travel Time	-1.97 (0.00)	-5.55 (0.00)	-4.61 (0.00)	-3.02 (0.00)	-5.54 (0.00)	-0.096 (0.00)	-0.0810 (0.00)
Travel Cost	-0.605 (0.00)	-0.619 (0.00)	-3.38 (0.00)	-1.45 (0.00)	-3.53 (0.00)	-7.92 (0.00)	-1.61 (0.00)

Values within parentheses are p-values.

## REFERENCES

- Abdulhai, B., Pringle, R., & Karakoulas, G. J. (2003). Reinforcement Learning for *True Adaptive Traffic Signal Control*. *Journal of Transportation Engineering*, 129(3), 278–285. [https://doi.org/10.1061/\(ASCE\)0733-947X\(2003\)129:3\(278\)](https://doi.org/10.1061/(ASCE)0733-947X(2003)129:3(278))
- Abou-Zeid, M., & Ben-Akiva, M. (2014). Hybrid choice models. In *Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar, Cheltenham*. (pp. 383–412). <https://doi.org/10.4337/9781781003152.00025>
- Aboutaleb, Y. M., Danaf, M., Xie, Y., & Ben-Akiva, M. (2021). Discrete Choice Analysis with Machine Learning Capabilities. *ArXiv Preprint ArXiv: 2101.10261v1*, 1–14. <http://arxiv.org/abs/2101.10261>
- Abramowitz, M., & Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.
- Al-Ayyash, Z., Abou-Zeid, M., & Kaysi, I. (2016). Modeling the demand for a shared-ride taxi service: An application to an organization-based context. *Transport Policy*, 48, 169–182. <https://doi.org/10.1016/j.tranpol.2016.02.013>
- Andrade, K., Uchida, K., & Kagaya, S. (2006). Development of Transport Mode Choice Model by Using Adaptive Neuro-Fuzzy Inference System. *Transportation Research Record, 1977*, 8–16. <https://doi.org/10.3141/1977-04>
- Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research*, 39(4), 479–487. <https://doi.org/10.1509/jmkr.39.4.479.19124>
- AUB Fact Book 2016-2017. (2016). In *Office of Institutional Research and Assessment (OIRA)*. [https://www.aub.edu.lb/oira/Documents/Fact Book/FB201617.pdf](https://www.aub.edu.lb/oira/Documents/Fact%20Book/FB201617.pdf)
- Bansal, P., Krueger, R., Bierlaire, M., Daziano, R. A., & Rashidi, T. H. (2020). Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. *Transportation Research Part B: Methodological*, 131, 124–142. <https://doi.org/10.1016/j.trb.2019.12.001>
- Ben-Akiva, M., & Abou-Zeid, M. (2013). Methodological issues in modelling time-of-travel preferences. *Transportmetrica A: Transport Science*, 9(9), 846–859. <https://doi.org/10.1080/18128602.2012.686532>
- Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis. Theory and Applications to Travel Demand*. MIT Press, Cambridge, Massachusetts.
- 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., de Palma, A., Gopinath, D., Karlstrom, A., & Munizaga, M. a. (2002). Hybrid Choice Models : Progress and Challenges Massachusetts Institute of Technology. *Marketing Letters*, 13(3), 163–175.
- Ben-Akiva, M., Walker, J., Bernardino, A., Gopinath, D., Morikawa, T., & Polydoropoulou, A. (2002). Integration of Choice and Latent Variable Model. In *Perpetual Motion: Travel Behaviour Research Opportunities and Application*

- Challenges*, H. Mahmassani (ed.) (pp. 431–470).
- Bhat, C. R. (1997). Endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science*, 31(1), 34–48.  
<https://doi.org/10.1287/trsc.31.1.34>
- Bhat, C. R. (1998). Accommodating Variations in Responsiveness to Level-of-Service Measures in Travel Mode Choice Modeling. *Transportation Research Part A: Policy and Practice*, 32(7), 495–507. [https://doi.org/10.1016/S0965-8564\(98\)00011-1](https://doi.org/10.1016/S0965-8564(98)00011-1)
- Bhat, C. R. (2000). Incorporating Observed and Unobserved Heterogeneity in Urban Work Travel Mode Choice Modeling. *Transportation Science*, 34(2), 228–238.  
<https://doi.org/10.1287/trsc.34.2.228.12306>
- Bhat, C. R., & Eluru, N. (2009). A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. *Transportation Research Part B: Methodological*, 43(7), 749–765.  
<https://doi.org/10.1016/j.trb.2009.02.001>
- Bhat, C. R., & Lawton, T. K. (2000). Passenger travel demand forecasting. *Transportation in the New Millennium*, 1–6.  
<https://doi.org/10.1177/0308275X9501500302>
- Biagioni, J. P., Szczurek, P. M., Nelson, P. C., & Mohammadian, A. (2009). Tour-Based Mode Choice Modeling : Using An Ensemble of ( Un- ) Conditional Data-Mining Classifiers. *Transportation Research Board 88th Annual Meeting*.
- Bierlaire, M. (2018). *Swissmetro*. URL: [https://transport.epfl.ch/documents/technicalReports/CS\\_SwissmetroDescription.pdf](https://transport.epfl.ch/documents/technicalReports/CS_SwissmetroDescription.pdf).  
[https://transport.epfl.ch/documents/technicalReports/CS\\_SwissmetroDescription.pdf](https://transport.epfl.ch/documents/technicalReports/CS_SwissmetroDescription.pdf)
- Bierlaire, M., Axhausen, K. W., & Abay, G. (2001). The acceptance of modal innovation: The case of Swissmetro. *1st Swiss Transport Research Conference*.
- Bierlaire, M., & Lurkin, V. (2017). Introduction to Disaggregate Demand Models. *INFORMS TutORials in Operations Research, October*, 48–67.
- Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a Mixture Model for Clustering with Integrated Completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 19–25.  
<https://doi.org/10.1109/TPAMI.2012.125>
- Bingham, E. (2001). Reinforcement Learning in Neurofuzzy Traffic Signal Control. *European Journal of Operational Research*, 131(2), 232–241.  
[https://doi.org/10.1016/S0377-2217\(00\)00123-5](https://doi.org/10.1016/S0377-2217(00)00123-5)
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.  
<http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop - Pattern Recognition And Machine Learning - Springer 2006.pdf>
- Brathwaite, T., Vij, A., & Walker, J. L. (2017). *Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice*.  
<https://doi.org/10.1086/664709>

- Bujosa, A., Riera, A., & Hicks, R. L. (2010). Combining Discrete and Continuous Representations of Preference Heterogeneity: A Latent Class Approach. *Environmental and Resource Economics*, 47(4), 477–493. <https://doi.org/10.1007/s10640-010-9389-y>
- Cantarella, G. E., & de Luca, S. (2005). Multilayer Feedforward Networks for Transportation Mode Choice Analysis: An analysis and a Comparison with Random Utility Models. *Transportation Research Part C: Emerging Technologies*, 13(2), 121–155. <https://doi.org/10.1016/j.trc.2005.04.002>
- Cherchi, E., & de Dios Ortúzar, J. (2006). On fitting mode specific constants in the presence of new options in RP/SP models. *Transportation Research Part A: Policy and Practice*, 40(1), 1–18. <https://doi.org/10.1016/j.tra.2005.04.002>
- Chorus, C. (2012). Random Regret Minimization: An Overview of Model Properties and Empirical Evidence. *Transport Reviews*, 32(1), 75–92. <https://doi.org/10.1080/01441647.2011.609947>
- Chorus, C., Arentze, T. A., & Timmermans, H. J. P. (2008). A Random Regret-Minimization model of travel choice. *Transportation Research Part B: Methodological*, 42(1), 1–18. <https://doi.org/10.1016/j.trb.2007.05.004>
- Cox, D. R. (1966). Some Procedures Connected with the Logistic Qualitative Response Curve. In: David, F.N. (Ed.), *Research Papers in Statistics*, John Wiley & Sons, New York. Pp. 55-71.
- Dabiri, S., & Heaslip, K. (2018). Inferring Transportation Modes from GPS Trajectories Using a Convolutional Neural Network. *Transportation Research Part C: Emerging Technologies*, 86(November 2017), 360–371. <https://doi.org/10.1016/j.trc.2017.11.021>
- Danaf, M., Becker, F., Song, X., Atasoy, B., & Ben-Akiva, M. (2019). Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems*, 119, 35–45. <https://doi.org/10.1016/j.dss.2019.02.003>
- Darnton, A. (2008). Practical Guide: An overview of behaviour change models and their uses. In *Government Social Research (GSR)* (Issue July). [http://resources.civilservice.gov.uk/wp-content/uploads/2011/09/Behaviour-change\\_practical\\_guide\\_tcm6-9696.pdf](http://resources.civilservice.gov.uk/wp-content/uploads/2011/09/Behaviour-change_practical_guide_tcm6-9696.pdf)
- Dekker, T., & Chorus, C. (2018). Consumer surplus for random regret minimisation models. *Journal of Environmental Economics and Policy*, 7(3), 269–286. <https://doi.org/10.1080/21606544.2018.1424039>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodologica)*, 39(1), 1–38. <https://doi.org/10.1177/019262339101900314>
- Deshpande, M., & Bajaj, P. R. (2017). Short Term Traffic Flow Prediction Based on Neuro-Fuzzy Hybrid Sytem. *Proceedings of 2016 International Conference on ICT in Business, Industry, and Government, ICTBIG 2016*, 1–3. <https://doi.org/10.1109/ICTBIG.2016.7892699>
- Ding, C., Cao, X., & Wang, Y. (2018). Synergistic effects of the built environment and

- commuting programs on commute mode choice. *Transportation Research Part A: Policy and Practice*, 118(August), 104–118. <https://doi.org/10.1016/j.tra.2018.08.041>
- Doshi-velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning. ML*, 1–13.
- El Zarwi, F. (2017a). *lccm, a Python package for estimating latent class choice models using the Expectation Maximization (EM) algorithm to maximize the likelihood function*. <https://pypi.org/project/lccm/0.1.20/>
- El Zarwi, F. (2017b). *Modeling and Forecasting the Impact of Major Technological and Infrastructural Changes on Travel Demand*. University of California, Berkeley (PhD Thesis).
- Errampalli, M., Okushima, M., & Akiyama, T. (2007). Combined fuzzy logic based mode choice and microscopic simulation model for transport policy evaluation. *11th World Conference on Transport Research*.
- Fosgerau, M., & Hess, S. (2009). Competing methods for representing random taste heterogeneity in discrete choice models. *European Transport*, 42, 1–25.
- Gammelli, D., Rolsted, K. P., Pacino, D., & Rodrigues, F. (2020). Generalized Multi-Output Gaussian Process Censored Regression. *ArXiv Preprint ArXiv:2009.04822*.
- Gazder, U., & Ratrou, N. T. (2016). A New Logit-Artificial Neural Network Ensemble for Mode Choice Modeling: A Case Study for Border Transport. *Journal of Advanced Transportation*, 49, 885–866. <https://doi.org/10.1002/atr>
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), 452–459.
- Gibbs, M. N., & Mackay, D. J. C. (2000). Variational Gaussian Process Classifiers. *IEEE Transactions on Neural Networks*, 11(6), 1458–1464.
- Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2014). Forecasting the demand for electric vehicles: Accounting for attitudes and perceptions. *Transportation Science*, 48(4), 483–499. <https://doi.org/10.1287/trsc.2013.0487>
- Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., & Mohammadian, A. (2018). Modeling Travel Mode and Timing Decisions: Comparison of Artificial Neural Networks and Copula-Based Joint Model. *Travel Behaviour and Society*, 10(October 2016), 21–32. <https://doi.org/10.1016/j.tbs.2017.09.003>
- Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L., & Perez, R. (2010). Automating Mode Detection for Travel Behaviour Analysis by Using Global Positioning Systems-Enabled Mobile Phones and Neural Networks. *IET Intelligent Transport Systems*, 4(1), 37. <https://doi.org/10.1049/iet-its.2009.0029>
- Gopinath, D. A. (1995). *Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand*.
- Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8), 681–698. [https://doi.org/10.1016/S0191-2615\(02\)00046-2](https://doi.org/10.1016/S0191-2615(02)00046-2)

- Greene, W. H., & Hensher, D. A. (2013). Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45(14), 1897–1902. <https://doi.org/10.1080/00036846.2011.650325>
- Greene, W. H., Hensher, D. A., & Rose, J. (2006). Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. *Transportation Research Part B: Methodological*, 40(1), 75–92. <https://doi.org/10.1016/j.trb.2005.01.005>
- Guevara, C. A. (2017). Mode-valued differences of in-vehicle travel time Savings. *Transportation*, 44(5), 977–997. <https://doi.org/10.1007/s11116-016-9689-3>
- Hagenauer, J., & Helbich, M. (2017). A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications*, 78, 273–282. <https://doi.org/10.1016/j.eswa.2017.01.057>
- Han, Y. (2019). *Neural-Embedded Discrete Choice Models*. Massachusetts Institute of Technology (PhD Thesis).
- Han, Y., Zegras, C., Pereira, F. C., & Ben-Akiva, M. (2020). A neural-embedded choice model: TasteNet-MNL modeling taste heterogeneity with flexibility and interpretability. *ArXiv Preprint ArXiv: 2002.00992v1*.
- Hensher, D. A., & Ton, T. T. (2000). A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice. *Transportation Research Part E: Logistics and Transportation Review*, 36(3), 155–172. [https://doi.org/10.1016/S1366-5545\(99\)00030-7](https://doi.org/10.1016/S1366-5545(99)00030-7)
- Hess, S., Ben-Akiva, M., & Gopinath, D. A. (2009). Taste heterogeneity, correlation, and elasticities in latent class choice models. *Compendium of Papers, 88th Annual Meeting of the Transportation Research Board. Transportation Research Board*.
- Hess, S., Bierlaire, M., & Polak, J. W. (2005). Estimation of Value of Travel-Time Savings Using Mixed Logit Models. *Transportation Research Part A: Policy and Practice*, 39(2-3 SPEC. ISS.), 221–236. <https://doi.org/10.1016/j.tra.2004.09.007>
- Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). Weak teachers: Assisted specification of discrete choice models using ensemble learning. *HEART 2019, 8th Symposium of the European Association for Research in Transportation, MI*, 1–12.
- Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., & Jin, Y. (2020). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, 38. <https://doi.org/10.1016/j.jocm.2020.100221>
- Hillel, T., Elshafie, M. Z. E. B., & Jin, Y. (2018). Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, 171(1), 29–42. <https://doi.org/10.1680/jsmic.17.00018>
- Hong, W. C., Dong, Y., Zheng, F., & Lai, C. Y. (2011). Forecasting Urban Traffic Flow by SVR with Continuous ACO. *Applied Mathematical Modelling*, 35(3), 1282–1291. <https://doi.org/10.1016/j.apm.2010.09.005>
- Idé, T., & Kato, S. (2009). Travel-Time Prediction using Gaussian Process Regression: A Trajectory-Based Approach. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 1185–1196.



- Jahangiri, A., & Rakha, H. A. (2015). Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), 2406–2417.
- Jahangiri, A., & Rakha, H. A. (2014). Developing a Support Vector Machine (SVM) Classifier for Transportation Mode Identification by Using Mobile Phone Sensor Data. *Transportation Research Board 93rd Annual Meeting, January*.
- Jin, X., Cheu, R. L., & Srinivasan, D. (2002). Development and adaptation of constructive probabilistic neural network in freeway incident detection. *Transportation Research Part C: Emerging Technologies*, 10(2), 121–147. <https://doi.org/10.1177/009365028601300102>
- Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. *Transportation Research Part C: Emerging Technologies*, 19(3), 387–399. <https://doi.org/10.1016/j.trc.2010.10.004>
- Keane, M., & Wasi, N. (2013). Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics*, 28(6), 1018–1045. <https://doi.org/10.1002/jae>
- Kedia, A. S., Saw, K. B., & Katti, B. K. (2015). Fuzzy logic approach in mode choice modelling for education trips: A case study of Indian metropolitan city. *Transport*, 30(3), 286–293. <https://doi.org/10.3846/16484142.2015.1081279>
- Kopitar, L., Cilar, L., Kocbek, P., & Stiglic, G. (2019). Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening. In M. Marcos, J. M. Juarez, R. Lenz, G. J. Nalepa, S. Nowaczyk, M. Peleg, J. Stefanowski, & G. Stiglic (Eds.), *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems* (pp. 108–119). Springer Nature. <https://doi.org/10.1007/978-3-030-37446-4>
- Krueger, R., Vij, A., & Rashidi, T. H. (2018). A Dirichlet Process Mixture Model of Discrete Choice. *ArXiv Preprint ArXiv:1801.06296*. <http://arxiv.org/abs/1801.06296>
- Kumar, M., Sarkar, P., & Madhu, E. (2013). Development of Fuzzy Logic Based Mode Choice Model Considering Various Public Transport Policy Options. *International Journal for Traffic and Transport Engineering*, 3(4), 408–425. [https://doi.org/10.7708/ijtte.2013.3\(4\).05](https://doi.org/10.7708/ijtte.2013.3(4).05)
- Lee, D., Derrible, S., & Pereira, F. C. (2018). *Comparison of Four Types of Artificial Neural Networks and a Multinomial Logit Model for Travel Mode Choice Modeling*. <https://doi.org/10.1177/0361198118796971>
- Liang, L., Xu, M., Grant-Muller, S., & Mussone, L. (2018). Travel Mode Choice Analysis Based on Household Mobility Survey Data on Milan: Comparison of the Multinomial Logit Model and Random Forest Approach. *Transportation Research Board 98th Annual Meeting*. <https://doi.org/10.1007/s12671-013-0199-5>
- Liu, S., Yue, Y., & Krishnan, R. (2013). *Adaptive Collective Routing Using Gaussian Process Dynamic Congestion Models*. 704–712.
- Liu, Y., Bansal, P., Daziano, R., & Samaranayake, S. (2019). A framework to integrate

- mode choice in the design of mobility-on-demand systems. *Transportation Research Part C: Emerging Technologies*, 105(August 2018), 648–665. <https://doi.org/10.1016/j.trc.2018.09.022>
- Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., & Ghahramani, Z. (2014). Automatic Construction and Natural-Language Description of Nonparametric Regression Models. *ArXiv:1402.4304*.
- Ma, D., Sheng, B., Jin, S., Ma, X., & Gao, P. (2018). Short-Term Traffic Flow Forecasting by Selecting Appropriate Predictions Based on Pattern Matching. *IEEE Access*, 6, 75629–75638. <https://doi.org/10.1109/ACCESS.2018.2879055>
- Mackay, D. J. C. (1997). Gaussian Processes: A Replacement for Supervised Neural Networks? In *Lecture notes for a tutorial in advances in neural information processing systems*.
- Mackay, D. J. C. (1998). Introduction to Gaussian Processes. In C. M. Bishop (Ed.), *Neural Networks and Machine Learning*. Springer-Verlag.
- Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK.
- Manski, C. F. (2001). Daniel McFadden and the Econometric Analysis of Discrete Choice. *Scandinavian Journal of Economics*, 103(2), 217–230. <https://doi.org/10.1111/1467-9442.00241>
- Manski, C. F. (2013). *Public policy in an uncertain world: analysis and decisions*. Harvard University Press.
- Marschak, J. (1960). *Binary Choice Constraints on Random Utility Indicators*. <https://econpapers.repec.org/paper/cwlcwldpp/74.htm>
- McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In: *Zarembka P. (Ed.), Frontiers in Econometrics*, New York: Academic Press, 105–142. <https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf>
- McFadden, D. (1986). The Choice Theory Approach to Market Research. *Marketing Science*, 5(4), 275–297.
- McFadden, D. (2001). Economic Choices. *The American Economic Review*, 91, 351–378.
- McFadden, D., & Train, K. E. (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15(May), 447–470. [http://download.clib.psu.ac.th/datawebclib/e\\_resource/trial\\_database/WileyInterScienceCD/pdf/JAE/JAE\\_3.pdf](http://download.clib.psu.ac.th/datawebclib/e_resource/trial_database/WileyInterScienceCD/pdf/JAE/JAE_3.pdf)
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, 6(1), 355–378. <https://doi.org/10.1146/annurev-statistics-031017-100325>
- McNicholas, P. D., & Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, 26(21), 2705–2712. <https://doi.org/10.1093/bioinformatics/btq498>
- Minal, S., & Sekhar, C. R. (2014). Mode Choice Analysis: the Data, the Models and Future Ahead. *International Journal for Traffic and Transport Engineering*, 4(3),

- 269–285. [https://doi.org/10.7708/ijtte.2014.4\(3\).03](https://doi.org/10.7708/ijtte.2014.4(3).03)
- Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference* (Vol. 4). Massachusetts Institute of Technology (PhD Thesis).
- Mohammadian, A., & Miller, E. J. (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(1), 92–100. <https://doi.org/10.3141/1807-12>
- Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. <https://doi.org/10.1016/j.dsp.2017.10.011>
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer-Verlag.
- Neal, R. M. (1999). Regression and Classification Using Gaussian Process Priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 6* (pp. 475–501). Oxford University Press.
- Nijkamp, P., Reggiani, A., & Trirapepe, T. (1996). Modelling Inter-Urban Transport Flows In Italy: A Comparison Between Neural Network Analysis And Logit Analysis.pdf. *Transportation Research Part C: Emerging Technologies*, 4(6), 323–338.
- Nisbet, R., Gary, M., & Ken, Y. (2017). *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier Science & Technology.
- Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. Springer.
- Nocedal, J., Wright, S. J., & Robinson, S. M. (1999). *Numerical Optimization*. Springer.
- Omrani, H. (2015). Predicting Travel Mode of Individuals by Machine Learning. *Transportation Research Procedia*, 10(July), 840–849. <https://doi.org/10.1016/j.trpro.2015.09.037>
- Omrani, H., Charif, O., Gerber, P., Awasthi, A., & Trigano, P. (2013). Prediction of Individual Travel Mode with Evidential Neural Network Model. *Transportation Research Record: Journal of the Transportation Research Board*, 2399, 1–8. <https://doi.org/10.3141/2399-01>
- Opper, M., & Winther, O. (2000). Gaussian Processes for Classification: Mean-Field Algorithms. *Neural Computation*, 12(11), 2655–2684.
- Pedregosa, F., Varoquaux, G., Gramfort, V., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Pulugurta, S., Arun, A., & Errampalli, M. (2013). Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model. *Procedia - Social and Behavioral Sciences*, 104, 583–592. <https://doi.org/10.1016/j.sbspro.2013.11.152>
- Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*.

the MIT Press.

- Ratrou, N. T., Gazder, U., & Al-Madani, H. M. N. (2014). A review of mode choice modelling techniques for intra-city and border transport. *World Review of Intermodal Transportation Research*, 5(1), 39–58.  
<https://doi.org/10.1504/WRITR.2014.065055>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). “ Why Should I Trust You ?” Explaining the Predictions of Any Classifier. *ArXiv:1602.04938*, 1135–1144.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Model-Agnostic Interpretability of Machine Learning. *ArXiv:1606.05386*, *Whi*.
- Richter, P., & Toledano-Ayala, M. (2015). Revisiting gaussian process regression modeling for localization in wireless sensor networks. *Sensors*, 15(9), 22587–22615. <https://doi.org/10.3390/s150922587>
- Robin, Th, Antonini, G., Bierlaire, M., & Cruz, J. (2009). Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1), 36–56. <https://doi.org/10.1016/j.trb.2008.06.010>
- Robin, Thomas, & Bierlaire, M. (2012). Modeling investor behavior. *Journal of Choice Modelling*, 5(2), 98–130. [https://doi.org/10.1016/S1755-5345\(13\)70054-X](https://doi.org/10.1016/S1755-5345(13)70054-X)
- Rodrigues, F. (2020). Scaling Bayesian inference of mixed multinomial logit models to very large datasets. *ArXiv Preprint ArXiv: 2004.05426v1*.
- Rodrigues, F., Borysov, S. S., Ribeiro, B., Member, S., & Pereira, F. C. (2016). A Bayesian additive model for understanding public transport usage in special events. 39(11), 2113–2126.
- Rodrigues, F., Henrickson, K., & Pereira, F. C. (2019). Multi-output Gaussian processes for crowdsourced traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 20(2), 594–603.
- Rodrigues, F., Ortelli, N., Bierlaire, M., & Pereira, F. C. (2020). Bayesian Automatic Relevance Determination for Utility Function Specification in Discrete Choice Models. *IEEE Transactions on Intelligent Transportation Systems*.  
<https://doi.org/10.5379/urbani-izziv-en-2006-17-01-02-017>
- Rodrigues, F., & Pereira, F. C. (2018). Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data. *Transportation Research. Part C: Emerging Technologies*, 95(January), 636–651.
- Sarle, W. S. (1994). Neural Networks and Statistical Models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994*, 1–13.  
<https://doi.org/10.1.1.27.699>
- Sayed, T., & Razavi, A. (2000). Comparison of Neural and Conventional Approaches to Mode Choice Analysis. *Journal of Computing in Civil Engineering*, 14(1), 23–30.  
<https://doi.org/10.1177/0145445514543465>
- Seeger, M. (2004). Gaussian Processes for Machine Learning. *International Journal of Neural Systems*, 14(2), 69–106.
- Sekhar, C. R., Minal, & Madhu, E. (2016). Multimodal Choice Modeling Using Random Forest. *International Journal for Traffic and Transport Engineering*, 6(3),

356–367.

- Sfeir, G., Abou-Zeid, M., & Kaysi, I. (2020). Multivariate count data models for adoption of new transport modes in an organization-based context. *Transport Policy*, 91(March), 59–75. <https://doi.org/10.1016/j.tranpol.2020.03.014>
- Sifringer, B., Lurkin, V., & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140, 236–261. <https://doi.org/10.1016/j.trb.2020.08.006>
- Sifringer, B., Lurkin, V., & Alahi, A. (2018). Enhancing Discrete Choice Models with Neural Networks. *18th Swiss Transport Research Conference, May*, 1–3.
- Srinivasan, D., Jin, X., & Cheu, R. L. (2004). Evaluation of Adaptive Neural Network Models for Freeway Incident Detection. *IEEE Transactions on Intelligent Transportation Systems*, 5(1), 1–11. <https://doi.org/10.1109/TITS.2004.825084>
- Srinivasan, D., Member, S., Choy, M. C., & Cheu, R. L. (2006). *Neural Networks for Real-Time Traffic Signal Control*. 7(3), 261–272. <https://doi.org/10.1109/TITS.2006.874716>
- Stathopoulos, A., Dimitriou, L., & Tsekeris, T. (2008). Fuzzy Modeling Approach for Combined Forecasting of Urban Traffic Flow. *Computer-Aided Civil and Infrastructure Engineering*, 23(7), 521–535. <https://doi.org/10.1111/j.1467-8667.2008.00558.x>
- Stein, M. L. (1999). *Interpolation of Saptial Data*. Springer-Verlag, New York.
- Subba Rao, P. V., Sikdar, P. K., Krishna Rao, K. V., & Dhingra, S. L. (1998). Another insight into artificial neural networks through behavioural analysis of access mode choice. *Computers, Environment and Urban Systems*, 22(5), 485–496. [https://doi.org/10.1016/S0198-9715\(98\)00036-2](https://doi.org/10.1016/S0198-9715(98)00036-2)
- Tang, L., Xiong, C., & Zhang, L. (2015). Decision Tree Method for Modeling Travel Mode Switching in a Dynamic Behavioral Process. *Transportation Planning and Technology*, 38(8), 833–850. <https://doi.org/10.1080/03081060.2015.1079385>
- Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 567–574.
- Train, K. (2016). Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, 19, 40–53. <https://doi.org/10.1016/j.jocm.2016.07.004>
- Train, K. E. (2008). EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1), 40–69. [https://doi.org/10.1016/S1755-5345\(13\)70022-8](https://doi.org/10.1016/S1755-5345(13)70022-8)
- Train, K. E. (2009). Discrete Choice Methods with Simulation. In *Cambridge University Press, Cambridge*. [https://doi.org/10.1016/S0898-1221\(04\)90100-9](https://doi.org/10.1016/S0898-1221(04)90100-9)
- Train, K., & Sonnier, G. (2004). Mixed Logit with Bounded Distributions of Correlated Partworths. In *In: Scarpa, R., Alberini, A.(Eds.), Applications of Simulation Methods in Environmental and Resource Economies. Kluwer Academic Publishers, Boston, MA*.
- Van Cranenburgh, S., & Alwosheel, A. S. (2019). An Artificial Neural Network Based

- Approach to Investigate Travellers' Decision-Rules. *Transportation Research Part C*, 98(September 2018), 152–166. <https://doi.org/S0968090X18305230>
- Varian, H. R. (1993). *What Use is Economic Theory?* (Issue 87).
- Vij, A., Carrel, A., & Walker, J. L. (2013). Incorporating the Influence of Latent Modal Preferences on Travel Mode Choice Behavior. *Transportation Research Part A: Policy and Practice*, 54, 164–178. <https://doi.org/10.1016/j.tra.2013.07.008>
- Vij, A., & Krueger, R. (2017). Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions. *Transportation Research Part B: Methodological*, 106, 76–101. <https://doi.org/10.1016/j.trb.2017.10.013>
- Vij, A., & Walker, J. L. (2014). Preference endogeneity in discrete choice models. *Transportation Research Part B: Methodological*, 64, 90–105. <https://doi.org/10.1016/j.trb.2014.02.008>
- Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2008). Temporal Evolution of Short-Term Urban Traffic Flow: A Nonlinear Dynamics Approach. *Computer-Aided Civil and Infrastructure Engineering*, 23(7), 536–548. <https://doi.org/10.1111/j.1467-8667.2008.00554.x>
- Vythoulkas, P. C., & Koutsopoulos, H. N. (2003). Modeling Discrete Choice Behavior Using Concepts from Fuzzy Set Theory, Approximate Reasoning and Neural Networks. *Transportation Research Part C: Emerging Technologies*, 11(1), 51–73. [https://doi.org/10.1016/S0968-090X\(02\)00021-9](https://doi.org/10.1016/S0968-090X(02)00021-9)
- Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343. [https://doi.org/10.1016/S0165-4896\(02\)00023-9](https://doi.org/10.1016/S0165-4896(02)00023-9)
- Walker, J. L., & Li, J. (2007). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1), 77–101. <https://doi.org/10.1007/s10109-006-0030-0>
- Wang, F., & Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation Research Record: Journal of the Transportation Research Board*. <https://doi.org/10.1177/0361198118773556>
- Wang, S., Wang, Q., & Zhao, J. (2020). Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118(July), 1–22. <https://doi.org/10.1016/j.trc.2020.102701>
- Wang, S., & Zhao, J. (2019). An empirical study of using deep neural network to analyze travel mode choice with interpretable economic information. *Transportation Research Board 98th Annual Meeting*.
- Wang, W., Chen, S., & Qu, G. (2008). Incident Detection Algorithm Based on Partial Least Squares Regression. *Transportation Research Part C: Emerging Technologies*, 16(1), 54–70. <https://doi.org/10.1016/j.trc.2007.06.005>
- Williams, C. K. I., & Barber, D. (1998). Bayesian Classification with Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1342–1351.

- Wong, M., & Farooq, B. (2019). ResLogit: A residual neural network logit model. *ArXiv Preprint ArXiv:1912.10058*.
- Wong, M., Farooq, B., & Bilodeau, G. A. (2018). Discriminative Conditional Restricted Boltzmann Machine for Discrete Choice and Latent Variable Modelling. *Journal of Choice Modelling*, 29, 152–168. <https://doi.org/10.1016/j.jocm.2017.11.003>
- Xian-Yu, J.-C. (2011). Travel Mode Choice Analysis Using Support Vector Machines. *ICCTP: American Society of Civil Engineers*, 360–371.
- Xie, C., Lu, J., & Parkany, E. (2003). Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 1854(03), 50–61. <https://doi.org/10.3141/1854-06>
- Xie, Y., Zhao, K., Sun, Y., & Chen, D. (2010). Gaussian Processes for Short-Term Traffic Volume Forecasting. *Transportation Research Record*, 1, 69–78. <https://doi.org/10.3141/2165-08>
- Yuan, Y., You, W., Boyle, K. J., Yuan, Y., You, W., & Boyle, K. (2015). *A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model parametric and nonparametric mixing distributions for*.
- Zhang, Y., & Xie, Y. (2008). Travel Mode Choice Modeling with Support Vector Machines. *Transportation Research Record: Journal of the Transportation Research Board*, 2076(1), 141–150. <https://doi.org/10.3141/2076-16>
- Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4), 550–560.
- Zhu, X., Li, J., Liu, Z., & Yang, F. (2017). Learning Transportation Mode Choice for Context-Aware Services with Directed-Graph-Guided Fused Lasso from GPS Trajectory Data. *Proceedings - 2017 IEEE 24th International Conference on Web Services, ICWS 2017*, 692–699. <https://doi.org/10.1109/ICWS.2017.83>

