IMPROVED ESTIMATION OF FLEXIBLE LOGIT MODELS AND AN EXTENSION TO A MODEL WITH A T-DISTRIBUTED ERROR KERNEL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Prateek Bansal

August 2019

IMPROVED ESTIMATION OF FLEXIBLE LOGIT MODELS AND AN EXTENSION TO A MODEL WITH A T-DISTRIBUTED ERROR KERNEL

Prateek Bansal, Ph.D.

Cornell University 2019

Understanding various micro-decisions of travelers (e.g., choice of vehicles, travel modes, or destinations) is of utmost importance in travel demand modeling. After the first application of the multinomial logit model in the early 1970s, micro-econometric models to elicit these type of choices have evolved in mainly two ways: a) computationally efficient estimation (e.g., fast integral approximations); and b) behaviorally defensible models (e.g., modeling preference heterogeneity).

This dissertation contributes to both lines of behavior modeling research — whereas chapters one to three analyze and improve the computational efficiency of flexible logit models and the required approximation of high-dimensional integrals, chapter four derives the first multinomial response model with a t-distributed error kernel that accounts for *decision uncertainty* behavior of travelers. A summary of each chapter is provided below.

In *Chapter 1*, we extend the logit-mixed logit (LML) model, an advanced semi-parametric specification of preference heterogeneity, to a combination of fixed and random parameters. We show that the likelihood of the LML specification loses its special properties due to the inclusion of fixed parameters, leading to a much higher estimation time. In an empirical application about preferences for alternative fuel vehicles in China, estimation time increased by a factor of 20-40 when

introducing fixed parameters. Despite losses in computation efficiency, we show that the flexible LML could retrieve multimodal mixing distributions.

In *Chapter 2*, we derive, implement, and test minorization-maximization (MM) algorithms to estimate the semiparametric LML and mixture-of-normals multinomial logit (MON-MNL) models. In a Monte Carlo study and empirical application to estimate consumer's willingness to adopt electric motorcycles in Indonesia, we compare the maximum simulated likelihood estimator (MSLE) with the derived MM algorithms. Whereas in LML estimation MM is computationally noncompetitive with MSLE, it is a competitive replacement to MSLE for MON-MNL that obviates computation of complex analytical gradients.

In *Chapter 3*, we propose the application of a moment-based designed quadrature (DQ) method to approximate multi-dimensional integrals in MSLE of discrete choice models. The results of simulation study indicate that DQ is a potentially attractive alternative to quasi-Monte Carlo (QMC) because it requires fewer evaluations of the conditional likelihood (i.e., lower computation time) as compared to QMC methods, is as easy to implement, ensures positivity of weights, and can be created on any general polynomial spaces. Finally, we validate the performance of DQ on a case study to understand preferences for mobility-on-demand services in New York City.

In *Chapter 4*, we demonstrate that using a t-distributed error kernel in multinomial choice models helps in better predicting the preferences in class-imbalance datasets. This specification also implicitly accounts for the consumers' *decision uncertainty* behavior. Because of these statistical and behavioral advantages, we de-

rive the first multinomial response model with a t-distributed error kernel and extend this to a generalized continuous-multinomial (GCM) model. In the empirical study related to the adoption of electric vehicles (EVs), we observe that accounting for *decision uncertainty* behavior in the GCM model with t-distributed error kernel results into a higher willingness to pay for improving the EV attributes than those of a GCM model with a normally-distributed error kernel. These differences are relevant in making policies to expedite the market penetration of EVs.

BIOGRAPHICAL SKETCH

Having grown up in densely populated cities across India, Prateek have experienced the challenges that delays and overcrowding in rails pose on the day-to-day lives of people. This, in turn, motivated him to pursue research in Transportation Economics and Engineering.

Before joining Cornell, Prateek received his M.S. from The University of Texas at Austin in Transportation Engineering and B.Tech. from Indian Institute of Technology (IIT) Delhi in Civil Engineering. Over the years he has also collaborated with researchers across different settings (e.g., Canada, Chile, South Africa, Sweden, Taiwan, and the United States) on diverse transportation-related projects.

His pre-doctoral research experiences made him realize that system-level decisions can be better informed by eliciting and aggregating individual's travel preferences. This helped him in focusing his doctoral research on advancing the micro-econometric models of consumer choice.

His current research interests are inter-disciplinary and broadly revolve around: i) bridging machine learning and econometrics, i.e. deriving computationally-efficient estimators of econometric models using approximate Bayesian inference; ii) addressing commonly ignored endogeneity concerns in multinomial choice models using partial identification; iii) understanding the impact of discrete choice experiment designs on the modeling results.

To my parents and sister

ACKNOWLEDGEMENTS

I take this opportunity to thank my mentors, collaborators, funding agencies, friends, and family, whose guidance and/or unconditional support have helped me sail through my doctoral journey.

My PhD committee has been very supportive throughout these years. I cannot thank Dr Ricardo Daziano (committee chair) enough for believing in my ideas and providing optimal freedom and time to pursue them. I would cherish his prompt feedback on research papers and our discussions during lunch meetings. He ensured that I acquire all the required skills and experiences to become a well-rounded academic researcher – suggested right course sequences, supported collaborations through conference visits and university exchanges, and provided opportunities to teach graduate classes and to write grant proposals. In addition, advanced econometrics courses taught by Drs Shanjun Li and Thomas DiCiccio have played a crucial role in setting the required foundation to start my research early. I am particularly thankful to them for providing statistical insights to handle various research problems that I encountered during my PhD.

Among other mentors, I am obliged to Prof. Kenneth Train for constantly acknowledging, reviewing, and encouraging my ideas, to Prof. Michel Bierlaire and Drs Ricardo Hurtubia, Taha Rashidi, Samitha Samaranayake, Erick Guerra, and Martin Achtnicht for providing survey data and/or detailed assessment of my work, to Prof. Kara Kockelman and Dr Stephen Boyles for teaching me fundamentals of research and making me realize my abilities as a researcher, and to Profs. Geetam Tiwari, Mu Chen Chen and Robert Chapleau for providing me

with a platform to pursue my passion for transportation science during my undergraduate research. Additionally, semester exchange experiences at the University of California, Berkeley and the University of Santiago, Chile have offered me unique exposure to multicultural research environments. I am grateful to Prof. Joan Walker, Dr Angelo Guevara, and Dr Alejandro Tirachini for hosting me and providing with such opportunities. I have been privileged to collaborate with multiple researchers during my PhD. Working with Yang, Rico, Rubal, Naveen, Tomas, and Subodh on multi-disciplinary problems have helped me learning various dimensions of research such as remote collaborations and patiently progressing multiple projects simultaneously. Such collaborations also resulted in prolific research output, thanks to my efficient collaborators.

I also acknowledge financial support for my graduate research, conference travel, and semester exchanges from the National Science Foundation, the Cornell Graduate School, the International Institute of Forecasters, and the King Abdullah Petroleum Studies and Research Center. However, any errors that remain are my sole responsibility.

I am blessed to have a close-knit circle of friends who enabled me to sustain moving-forward attitude during these years. Naveen has always been available to discuss any kind of professional and personal challenges. His invaluable guidance has shaped various aspects of my personality. Without his support, I could not have achieved even three-fourths of what I did in PhD. Yogesh has always been a great source of inspiration and kept alive my belief in hard work. Being seniors in the same discipline, Tarun, Prasad, and Rahul specifically helped in

making various academic decisions. I am also thankful to other friends at Cornell – Abhi, Seema, Udit, Prankur, Gargi, Jessica, Vidya, Payal, Rohil, Mathew, Anaka, Karen, Nima, Raashid, Faisal, Will, and Sophia – who ensured a vibrant life besides work by organizing house parties, potlucks, and much more. Aditya, Praveen, Saqib, Surabhi, Oscar, Mengqiao, and Mostafa made Berkeley feel like home during my exchange. Esteban, Felipe, Kim, Diego, Javiera, Jacqueline, Isabelle, Nico, Raul, Alisson, and Hans put a collective effort to make my visit to Chile memorable by providing authentic Latino experiences – from teaching me Spanish to hosting at family house parties. I am grateful to have friends like Averi, Kratika, Nikit, Rajat, Sunay, Dheeraj, Rishabh, Prateek Raj, Kartik, TKS, and Rohit Prakash, who latently provided me mental strength during tough times. Above all, I am fortunate to have parents and sister who absorbed financial adversities, trusted my decisions, and gave me the freedom to carve out my own path. This dissertation is dedicated to their sacrifices and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

EXTENDING THE LOGIT-MIXED LOGIT MODEL FOR A COMBINATION OF RANDOM AND FIXED PARAMETERS

1.1 Introduction: flexible mixing distributions

Analysts who model decision-making processes using random utility maximization theory cannot possibly include all relevant factors that affect choices. In fact, the decision-making process is generally heterogeneous. Heterogeneity may occur due to taste variations in how decision makers weigh different attributes. Taste variations can be decomposed into observed and unobserved preference heterogeneity. Whereas the standard conditional or multinomial logit (MNL) model (McFadden, 1973) can model observed preference heteorgeneity typically by interacting attributes with characteristics of the individual, unobserved preference heterogeneity requires further assumptions and a different model. To take into account unobserved preference heterogeneity, Boyd and Mellman (1980) introduced the mixed multinomial logit (MMNL) model by adding to MNL random parameters that follow a prespecified parametric, continuous mixing distribution. MMNL rapidly became standard practice in choice modeling research after the seminal paper by McFadden and Train (2000), where the authors showed that any random utility maximization model can be approximated by MMNL, if mixing distributions of the random parameters are specified correctly.

In the MMNL literature, most studies have used normal or lognormal heterogeneity distributions and a few have used gamma or triangular mixing. However, Louviere and Eagle (2006), Fosgerau and Hess (2007), and Louviere and Meyer (2008) have argued that the normal (and other parametric) mixing distributions may introduce problems of misspecification if the assumed distribution is not appropriate for the data; for example, researchers may obtain a negative marginal utility of income if a normal distribution is assumed (or any parametric distribution with a possibly negative support). Bajari et al. (2007), Fosgerau and Bierlaire (2007), Train (2008), Bastin et al. (2010), Fox et al. (2011), and Fosgerau and Mabit (2013) have specified non-parametric mixing distributions that are flexible and yet results into computationally less expensive estimation breaking down the traditional flexibility versus ease of estimation tradeoff.

MMNL estimation is a nonlinear optimization problem, but Bajari et al. (2007) proposed a method that is fast and easy to code that takes advantage of a linear-regression-type specification. The authors assume that the population can be sorted into finite classes or clusters (i.e. the number of preference parameters is discrete, c.f. the latent class logit specification: Kamakura and Russell, 1989; DeSarbo et al., 1995; Bhat, 1997) and assert that their estimator is non-parametric because any mixing distribution can be approximated by making the number of classes large enough. However, this linear regression method may violate some necessary constraints on the model parameters. To handle this issue, Fox et al. (2011) raparameterized MMNL and derived a specification very similar to that of Bajari et al. (2007), but used inequality constrained linear least squares. Fos-

gerau and Bierlaire (2007) further proposed a method to approximate any continuous distribution using Legendre polynomials. The use of polynomials is a very flexible simply because different distributions can be recovered by adding more terms to the series expansion. In addition, Train (2008) used computation-efficient expectation-maximization (EM) algorithms for non-parametric estimation of random parameter logit-type models (cf. Bhat, 1997).

Train (2016) has very recently proposed the **logit-mixed logit** (LML) model, which is a generalized specification for all above four methods to generate mixing distributions.[1] The logarithm of the mixing distribution can be easily specified in LML, using splines, polynomials, step functions, and many other functional forms. Additionally, a computationally-convenient likelihood equation – that does not require computation of conditional choice probabilities in iterative optimization – significantly reduces estimation time of LML (see Section 1.2.3 for details). However, in its original formulation, LML assumes all utility parameters to be random. Note, however, that fixed parameters are usually needed in practice: a few of such instances are discussed below and illustrated in the empirical study. First, fixed alternative-specific constants (ASCs) should be included in the utility to account for alternative-specific fixed effects. Second, preferences

[1]However, note that Train (2016) does not generalize Bastin et al. (2010) and Fosgerau and Mabit (2013). Fosgerau and Mabit (2013) suggests to draw random numbers from some *initial distribution* (e.g., uniform) and transform these draws using a polynomial or other function to recover the mixing distribution. Similarly, Bastin et al. (2010) proposed a non-parametric method to approximate the inverse cumulative distribution function of the mixing distribution. They use a polynomial approximation (B-spline parameterization) of an initially chosen uniform distribution. A major limitation of these procedures is understanding the relationship between the shape of the mixing distribution and that of the *initial distributions*.

for a specific covariate may not vary across individuals and thus the estimated parameter may be just non-random. Third, a parsimonious specification of heterogeneity combines different means (deterministic taste variation) with the same variance of the unobserved part, by allowing the mean of a random parameter (e.g., marginal utility of price) to differ by socio-demographics (e.g., gender). This taste variation specification requires the inclusion of a fixed parameter on the interaction of the covariate with the demographic dummy (e.g., price × gender). This chapter asserts that under a more general utility specification, i.e. utility with a combination of random and fixed parameters, the likelihood equation loses its computationally-convenient form, and thus LML exhibits major losses in computational efficiency. In fact, the computation time of a model with a fixed parameter may be even 40 times higher than that of the case where all parameters are random (see Section 1.3.2 for an example of this situation). This result is somewhat counter-intuitive because in parametric MMNL models assuming fixed parameters actually reduces computation time.

In sum, this chapter aims at: 1. Extending the LML model for a combination of random and fixed parameters (Section 1.2); and 2. Providing an empirical application of our LML extension (Section 1.3) – as case study we use stated preferences for alternative fuel vehicles in China. The original LML and our LML extension are used to fit models, and we compare the results with MNL and MMNL with normal heterogeneity (MMNL-N) models.

1.2 Incorporating a subset of fixed parameters in logit-mixed logit models

1.2.1 Model Specification

Consider a standard discrete choice setting where individual $n \in \{1, \ldots, N\}$ chooses one alternative from the mutually exclusive choice set $\{1, \ldots, J\}$ (indexed by j) over the set of discrete time periods $\{1, \ldots, T\}$ or choice situations (indexed by t). The random utility maximization model is specified as

$$U_{njt} = \mathbf{x}_{njt}'\boldsymbol{\beta}_n + \varepsilon_{njt} = \left[\mathbf{x}_{njt}^{F\prime}, \mathbf{x}_{njt}^{R\prime}\right]\begin{bmatrix} \boldsymbol{\beta}^F \\ \boldsymbol{\beta}_n^R \end{bmatrix} + \varepsilon_{njt} \tag{1.1}$$

where U_{njt} is the random indirect utility associated with individual n choosing alternative j during choice situation t, and ε_{njt} is an iid extreme value type I preference shock. Moreover, both the alternative attributes and preference parameters are sorted in two groups. On the one hand, $\boldsymbol{\beta}^F$ is a vector of fixed preference parameters and $\mathbf{x}_{njt}^F$ is the attribute/covariate vector associated with these fixed parameters. On the other hand, $\boldsymbol{\beta}_n^R$ is a vector of random parameters and $\mathbf{x}_{njt}^R$ is the attribute vector for which the researcher expects the presence of unobserved preference heterogeneity. The consideration of a combination of fixed and random parameters is a generalization of the LML model as derived by Train (2016), where all parameters are assumed random. The mixing distribution of the set of random parameters $\boldsymbol{\beta}_n^R$ is modeled semi-parametrically below.

If i_{nt} denotes the alternative observed to be chosen by individual n at time t, consider now the sequence of chosen alternatives for the decision maker $\{i_{n1}, \ldots, i_{nT}\}$. The probability that individual n made this sequence of choices, conditional on $\boldsymbol{\beta}_n$, is:

$$L_n(\boldsymbol{\beta}_n) = \prod_{t=1}^{T} Q_{i_{nt}t}(\boldsymbol{\beta}_n) \tag{1.2}$$

where $Q_{ni_{nt}t}(\boldsymbol{\beta}_n)$ is the probability of individual n choosing alternative i_{nt} in choice situation t. The conditional choice probability $Q_{ni_{nt}t}(\boldsymbol{\beta}_n)$ is given by the following conditional logit expression:

$$Q_{ni_{nt}t}(\boldsymbol{\beta}_n) = \frac{e^{U_{ni_{nt}t}}}{\sum_{j=1}^{J} e^{U_{njt}}}. \tag{1.3}$$

Variations in the set of random parameters $\boldsymbol{\beta}_n^R$ are represented semi-parametrically with a discrete mixing distribution over a finite support set S. Consider the following logit-type expression for the probability that $\boldsymbol{\beta}_n^R = \boldsymbol{\beta}_r^R$:

$$w_n(\boldsymbol{\beta}_r^R|\boldsymbol{\alpha}) = \Pr(\boldsymbol{\beta}_n^R = \boldsymbol{\beta}_r^R) = \frac{e^{\mathbf{z}(\boldsymbol{\beta}_r^R)'\boldsymbol{\alpha}}}{\sum_{s\in S} e^{\mathbf{z}(\boldsymbol{\beta}_s^R)'\boldsymbol{\alpha}}} \tag{1.4}$$

where $\boldsymbol{\alpha}$ is a vector of parameters and $\mathbf{z}(\boldsymbol{\beta}_r^R)$ is a vector-valued function that captures the shape of the mixing distribution. $\mathbf{z}$ can be specified as a sieve function, such as polynomial or other functional forms, including step functions and splines (see details in Train, 2016).

The unconditional probability of the sequence of choices of individual n (P_n) is simply:

$$P_n(\boldsymbol{\beta}^F, \boldsymbol{\alpha}) = \sum_{r\in S} L_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R) w_n(\boldsymbol{\beta}_r^R|\boldsymbol{\alpha}), \tag{1.5}$$

where the parameters of interest are $\boldsymbol{\beta}^F$ and $\boldsymbol{\alpha}$.[2] It is important to note that in the LML model as derived by Train (2016), i.e. with all parameters random, $\boldsymbol{\alpha}$ is the only parameter of interest. In our extension that incorporates fixed parameters, both $\boldsymbol{\beta}^F$ and $\boldsymbol{\alpha}$ need to be estimated. In what follows, we will label our extension to the model LML-FR, to make evident the combination of fixed and random parameters.

1.2.2 Maximum Likelihood Estimator

Adopting a frequentist approach to the estimation of the parameters of interest, the maximum likelihood estimator is implemented. The loglikelihood of the LML-FR model is shown in equation 1.6:

$$\mathcal{L}(\boldsymbol{\beta}^F, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \ln\Big(\sum_{r \in S} L_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R) w_n(\boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) \Big). \tag{1.6}$$

As pointed out by Train (2016), the discrete support set S of the mixing distribution may be too large for practical evaluation of the loglikelihood. Using simulation-based econometrics, the loglikelihood can be simulated by the standard procedure of sampling the random parameters. In the LML model, the loglikelihood is simulated by considering an individual-specific, randomly generated subset S_n of the original support S. The simulated loglikelihood can be then writ-

[2] The random parameters $\boldsymbol{\beta}_n^R$ are fully represented by the exogenous choice of $\mathbf{z}$ and the estimates of $\boldsymbol{\alpha}$.

ten as:

$$\tilde{\mathcal{L}}(\boldsymbol{\beta}^F, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \ln\Big(\sum_{r \in S_n} L_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R) w_n(\boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) \Big). \tag{1.7}$$

The partial derivative of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{\alpha}$ is:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\alpha}} = \sum_{n=1}^{N} \sum_{r \in S_n} \Big(h_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) - w_n(\boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) \Big) \mathbf{z}(\boldsymbol{\beta}_r^R) \tag{1.8}$$

where

$$h_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) = \frac{L_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R) w_n(\boldsymbol{\beta}_r^R | \boldsymbol{\alpha})}{\sum_{s \in S_n} L_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_s^R) w_n(\boldsymbol{\beta}_s^R | \boldsymbol{\alpha})}; \tag{1.9}$$

and the partial derivative of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{\beta}^F$ is:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\beta}^F} = \sum_{n=1}^{N} \sum_{r \in S_n} \left(h_n(\boldsymbol{\beta}^F, \boldsymbol{\beta}_r^R | \boldsymbol{\alpha}) \sum_{t=1}^{T} \left(\mathbf{x}_{ni_{nt}t}^F - \sum_{j=1}^{J} \mathbf{x}_{njt}^F Q_{njt} \right) \right). \tag{1.10}$$

Finally, the simulated score (gradient of $\tilde{\mathcal{L}}$) is:

$$\nabla(\tilde{\mathcal{L}}) = \left[\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\alpha}} \quad \frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\beta}^F} \right]. \tag{1.11}$$

The willingness-to-pay space specification is presented in Appendix A.1.

1.2.3 Computational Efficiency of LML with all random and some fixed parameters

If all parameters are random, i.e. $\boldsymbol{\beta}_n = \boldsymbol{\beta}_n^R$, the score of the model is simply $\partial \tilde{\mathcal{L}} / \partial \boldsymbol{\alpha}$ (equation 1.8). In addition, $L_n(\boldsymbol{\beta}_n^R)$ (equation 1.2, conditional on $\boldsymbol{\beta}_n$) is independent

of α and therefore, does not change in the optimization process. In the original LML model with all random parameters (Train, 2016), $L_n(\beta_n^R)$ is only computed once –before starting the optimization process– considerably reducing estimation time. However, if both random and fixed parameters are considered (LML-FR), the individual probability of the sequence of choices $L_n(\beta^F, \beta_n^R)$ changes at each iteration of the maximum simulated likelihood estimator, due to iterative changes in β^F. These iterative changes cause considerable losses in computational efficiency of LML-FR.

The covariance matrix of the parameters α and β^F can be estimated using a sandwich estimator based on score functions (equations 1.8 and 1.10). However, the mean and standard deviation (and other statistics) of β_n^R are of actual interest to researchers (rather than α itself). The delta method can be used to compute standard error of these statistics from the covariance matrix of α, but this calculation requires deriving and coding complex derivatives for different functional forms of $\mathbf{z}(\beta_r^R)$. Thus, standard errors of statistics associated with β_n^R and β^F are calculated through bootstrapping (Train, 2016). The loss of computational efficiency in LML-FR also affects the speed of calculating standard errors.

1.3 Empirical Application

To illustrate the use of the LML model and its LML-FR extension derived in this chapter (with a combination of fixed and random parameters), we use choice mi-

crodata about preferences for alternative fuel vehicles in China. Note that we only consider z to be a polynomial and spline to retrieve the mixing distribution of $\boldsymbol{\beta}_n^R$ (for further details about implementation of these functions in the estimator, see Train, 2016).

1.3.1 Data Description

The data comes from a discrete choice experiment (DCE) on alternative-fuel vehicles that was part of a larger survey of Chinese urban households. The survey was administered between December 2012 and January 2013 by Horizon Research Consultancy Group, a well-established market research company from China, using computer-assisted in-person interviews. Respondents aged 18 to 60 years were randomly intercepted and interviewed at malls or other public places of Bejing and Guangzhou. The final sample comprises 578 individuals.

In the DCE, respondents had the choice of four vehicles with different fuel types: gasoline, natural gas, hybrid, and all-electric. In addition to fuel type, the vehicles were described by five four-level attributes, as shown in Table 1.1. The levels of purchase price and engine power were customized based on lower and upper bounds that respondents indicated for the two attributes when questioned about their expected next vehicle purchase. The midpoint of the range spanned by the bound values served as individual reference, with the help of which the attribute levels of the underlying experimental design, expressed in relative terms,

were re-expressed in absolute terms. Fuel availability was represented by the share of existing filling stations that have the fuel, just as in previous studies in Europe (e.g., Achtnicht et al., 2012) and the U.S. (e.g., Bunch et al., 1993). The two lowest fuel availability levels were excluded for gasoline and hybrid cars in order to avoid unrealistic attribute-level combinations. Likewise, the lowest emission level was only applied to electric vehicles.

Table 1.1: Attributes and Levels for the Discrete Choice Experiment

Attribute	**Levels**
Fuel type	Gasoline, hybrid, natural gas, electric
Purchase price	70%, 90%, 110%, 130% of reference (in CN¥)[a]
Engine power	70%, 90%, 110%, 130% of reference (in kw)[a]
Fuel costs per 100 km	CN¥ 20, CN¥ 60, CN¥ 100, CN¥ 140
Emission level	10%[b], 50%, 100%, 150% of a present-day average vehicle
Fuel availability	10%[c], 40%[c], 70%, 100% of existing filling stations

[a] Midpoint of the range indicated by the respondent for the next purchase.
[b] Only applied to hybrid and electric vehicles.
[c] Only applied to natural gas and electric vehicles.

The final experimental design was generated with the help of Ngene software, using a D-efficient design that decreases the variance of parameter estimates (Kuhfeld et al., 1994). The thus generated 24 choice sets were divided into four blocks of six choice sets. Each respondent faced one of these choice set blocks, resulting in six observations per subject. In every choice set each fuel type appeared exactly once, where the order of the fuel types was randomized between choice sets. Respondents were asked to select the vehicle they preferred most and

consider their own budget constraints in their decision making.

1.3.2 Estimates

In addition to LML semi-parametric specifications, we also estimated MMNL-N models with a normal mixing distribution.[3] In all models, pairwise correlation of the parameters is assumed to be zero.

For LML specifications, the mixing distribution is represented by considering the functional form of $\mathbf{z}(\boldsymbol{\beta}_n^R)$ (see equation 1.4), either as a second-order polynomial (L) or a one-knot (K) spline. These LML specifications and MMNL-N result in the same number of identifiable parameters (G)[4]. For the final model specification, we also exploited the flexibility of LML and adopted a higher number of identifiable parameters, $\mathbf{z}(\boldsymbol{\beta}_n^R)$ to be a fourth order polynomial and a spline with three knots, to check the eventual presence of multimodality in the mixing distribution. If the number of random parameters is R and there are F fixed parameters, then the total number of identifiable parameters is: $G_{polynomial} = L \times R + F$, and $G_{spline} = (K + 1) \times R + F$. The MMNL-N estimates are used to set the grid of the LML models. The grid boundaries of the random parameters were set to two standard deviations away from the corresponding MMNL-N mean. Each dimension of the grid was divided in 1,000 equal-spaced points and 2,000 draws (S_n, see equation

[3] Results of a simple MNL model are also reported as baseline.

[4] To compare model fit, the Akaike Information Criterion (AIC) $= -2 \times \ln(SLL) + 2 \times G$ and the Bayesian Information Criterion (BIC) $= -2 \times \ln(SLL) + \ln(N) \times G$, where N is the number of observations, are used.

1.7) were drawn out from the multidimensional grid of $10^{3\times R}$ points (S). To derive standard errors, 100 bootstrap samples were used.

Since all the models are logit based, the estimates are comparable. MMNL-N was estimated in R using the gmnl package (Sarrias and Daziano, 2016). For LML estimation, the MATLAB code by (Train, 2016) was used. Note that the original code was written in willingness-to-pay space; however, we made modifications in the code to work in preference space. Furthermore, significant modifications to the code were implemented to allow for the inclusion of fixed parameters in the LML-FR specification. These major modifications include the appropriate computation of the likelihood function, its gradient (see equation 1.10), and $L_n(\boldsymbol{\beta}_n)$.

We first assumed all parameters to be random, and later incorporated fixed parameters when refining this specification. Table 1.2 shows the parameter estimates and Z-statistic of the first specification. A very low Z-statistic of the standard deviation of the marginal utility of engine power in MMNL-N-1 and variants of LML-1 support preference homogeneity for power across decision-makers in the sample. The standard deviation of the marginal utility of emission levels in LML-1 (Poly-order-2) is not statistically significant (see column 7 in Table 1.2), also providing evidence for further investigation.

As next steps, ASCs and fixed parameters are sequentially incorporated. The results of intermediate specifications are summarized in Table 1.3. While the mean

Table 1.2: Estimation Results - All Random Parameters

	MNL-1		MMNL-N-1		LML-1 (Poly-order-2)		LML-1 (Spline-knot-1)	
	Estimate	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**
Purchase Price	-0.398	-1.1	-0.533	-1.2	-0.559	-1.0	-0.505	-1.3
Engine Power	0.243	3.1	0.270	3.2	0.314	6.3	0.315	9.8
Fuel Costs	-0.213	-4.8	-0.267	-5.1	-0.262	-4.5	-0.261	-4.9
Emission Levels	0.088	2.2	0.129	2.9	0.123	3.1	0.123	2.7
Fuel Availability	0.732	10.5	1.004	9.3	0.991	9.2	0.995	9.8
SD.Purchase Price	-	-	2.417	2.6	2.216	2.0	2.211	2.3
SD.Engine Power	-	-	- 0.022	**0.1**	0.006	**0.4**	0.020	**0.9**
SD.Fuel Costs	-	-	0.487	4.5	0.539	3.1	0.540	3.7
SD.Emission Levels	-	-	0.228	1.5	0.021	**0.3**	0.134	1.5
SD.Fuel Availability	-	-	1.661	12.7	1.596	12.7	1.607	13.7
Mixing parameters	0		10		10		10	
Fixed parameters	5		0		0		0	
Loglikelihood	-4726.07		-4668.87		-4662.87		-4666.29	
AIC	9462.13		9357.74		9345.74		9352.58	
BIC	9492.89		9419.25		9407.25		9414.09	

of the marginal utility of emission levels is significant in MMNL-N-1[5], inclusion of ASCs in MNL-2 and MMNL-N-2 (see columns 2–5 in Table 1.3) results into statiscally insignificant estimates of emissions. MMNL-N-2 confirms the finding of MMNL-N-1 about homogeneity in preferences for engine power. Thus, specification 3 was obtained by eliminating emission levels and assuming a fixed marginal utility of engine power (see columns 6 to 11 in Table 1.3).

[5]In MNL-1 and MMNL-N-1 specification, the significant positive correlation between emissions and vehicle choice may be an artifact of not controlling for ASCs (fuel specific fixed-effects). Gasoline cars were most frequently selected (37%) and are on average dirtier than their hybrid and electric counterparts.

Table 1.3: Estimation Results - Inclusion of ASCs and Fixed Parameters

	MNL-2		MMNL-N-2		MMNL-N-3		LML-FR-3 (Poly-order-2)		LML-FR-3 (Spline-knot-1)	
	Estimate	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**
Gas:(intercept)	-0.091	-1.6	-0.101	-1.7	-0.094	-1.6	-0.097	-1.2	-0.097	-1.2
Gasoline:(intercept)	0.538	9.8	0.574	9.9	0.577	10.2	0.579	5.6	0.580	5.6
Hybrid:(intercept)	0.143	2.5	0.176	3.0	0.176	3.0	0.180	2.5	0.181	2.5
Purchase Price	-0.708	-1.9	-1.136	-2.3	-1.198	-2.4	-1.117	-2.1	-1.119	-2.2
Engine Power	0.195	2.5	0.195	2.3	0.184	2.2	0.193	2.6	0.193	2.6
Fuel Costs	-0.244	-5.4	-0.283	-5.3	-0.287	-5.4	-0.287	-5.3	-0.287	-5.5
Emission Level	0.002	**0.1**	0.041	**0.9**	-	-	-	-	-	-
Fuel Availability	0.277	3.3	0.511	4.3	0.504	4.3	0.526	4.9	0.525	4.8
SD.Purchase Price	-	-	3.315	3.6	3.356	3.7	2.775	3.5	2.878	4.0
SD.Engine Power	-	-	0.018	**0.1**	-	-	-	-	-	-
SD.Fuel Costs	-	-	0.491	4.6	0.494	4.6	0.484	4.4	0.489	4.1
SD.Emission Level	-	-	0.318	2.7	-	-	-	-	-	-
SD.Fuel Availability	-	-	1.669	12.5	1.636	12.6	1.679	13.9	1.685	14.1
Mixing parameters	0		10		6		6		6	
Fixed parameters	8		3		4		4		4	
Loglikelihood	-4639.44		-4581.30		-4581.79		-4580.87		-4580.89	
AIC	9294.88		9188.61		9183.57		9181.74		9181.78	
BIC	9344.09		9268.58		9245.08		9243.25		9243.29	

Table 1.4: Estimation Results - Final Specification

	MMNL-N-4		LML-FR-4 (Poly-order-2)		LML-FR-4 (Spline-knot-1)		LML-FR-4 (Poly-order-4)		LML-FR-4 (Spline-knot-3)	
	Estimate	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**	**Estimate**	**Z-stat**
Gas:(intercept)	-0.095	-1.6	-0.099	-1.2	-0.099	-1.2	-0.104	-1.2	-0.107	-1.3
Gasoline:(intercept)	0.576	10.2	0.579	5.7	0.580	5.7	0.566	5.3	0.551	5.4
Hybrid:(intercept)	0.175	3.0	0.181	2.5	0.182	2.5	0.171	2.3	0.154	2.1
Purchase Price	-1.199	-2.4	-1.127	-2.1	-1.130	-2.2	-1.220	-2.3	-1.489	-2.8
Engine Power	0.185	2.2	0.195	2.7	0.195	2.7	0.211	2.9	0.202	2.8
Fuel Costs	-0.287	-5.4	-0.286	-5.3	-0.287	-5.5	-0.292	-5.6	-0.243	-4.2
Fuel Availability	0.300	2.0	0.316	2.0	0.313	1.9	0.457	2.1	0.642	2.3
Fuel availability × male	0.416	2.0	0.461	2.5	0.466	2.4	0.359	2.1	0.250	2.3
SD.Purchase Price	3.379	3.7	2.829	3.4	2.929	3.9	3.255	3.7	3.379	5.5
SD.Fuel Costs	0.495	4.6	0.490	4.4	0.495	4.1	0.590	3.6	0.446	3.0
SD.Fuel Availability	1.621	12.5	1.711	13.1	1.719	13.2	1.813	14.6	1.802	13.6
Mixing parameters	6		6		6		12		12	
Fixed parameters	5		5		5		5		5	
Loglikelihood	-4579.71		-4578.29		-4578.31		-4571.95		-4567.71	
AIC	9181.42		9178.58		9178.62		9177.90		9169.42	
BIC	9249.08		9246.24		9246.28		9282.47		9273.99	

We further sequentially explored specifications where marginal utility of alternative-specific attributes vary with socio-demographics. Table 1.4 shows estimates of the final specification that indicates gender-based statistical differences in the mean of the marginal utility of fuel availability. The model fit statistics (log-likelihood, AIC, and BIC) and resultant estimates of MMNL-N-4 are close to that of LML-FR-4 variants with the same number of parameters (see columns 2-7 in Table 1.4).

The estimates indicate that the average Chinese car buyer prefers more powerful cars, but also likes to save on fuel costs. A dense fueling network is highly appreciated, as it guarantees the desired level of flexibility and mobility. As expected, the mean of the purchase price coefficient is negative but also turns out be statistical significant, unlike the first specification with all random parameters (Table 1.2).

The more flexible LML-FR-4 specifications, i.e. with more parameters (see columns 8–11 in Table 1.4), resulted into a higher loglikelihood at convergence but worse BIC values than that of specifications with a lower number of parameters. Thus, LML-FR with a lower number of parameters appears as preferable based on BIC.[6] Whereas the mean and standard deviation estimates was similar for both specifications, more flexible specifications were able to capture multimodality of the mixing distribution (see Figures 1.1, 1.2, and 1.3), which cannot be retrieved with a smaller number of parameters.

[6] Although BIC is a standard criterion in the literature to compare different models, the use of BIC can be sometimes misleading in deciding an appropriate model specification.

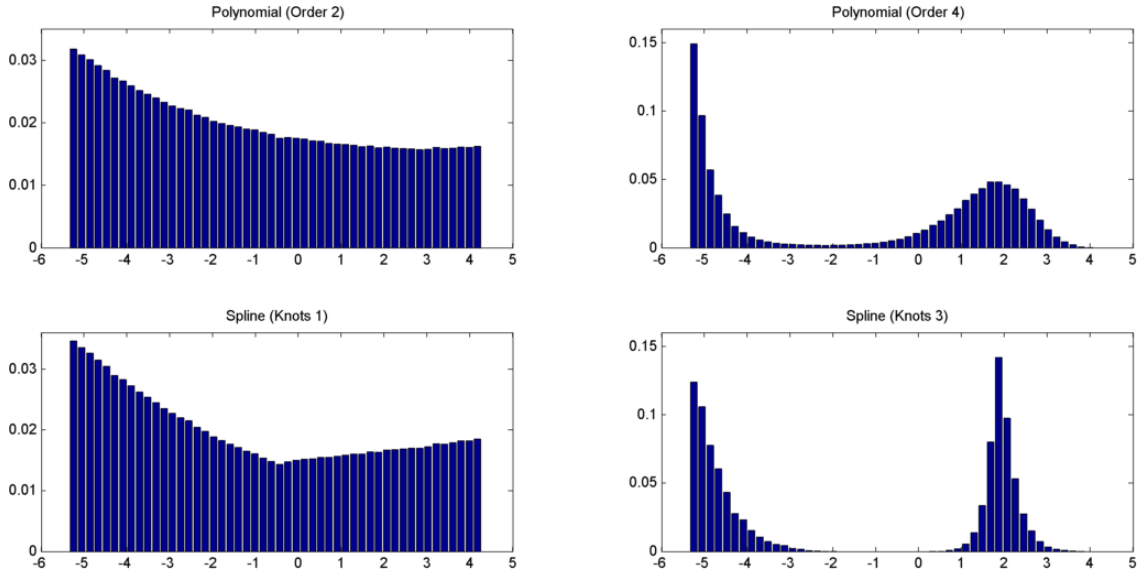

Figure 1.1: Histogram of Marginal Utility of Purchase Price

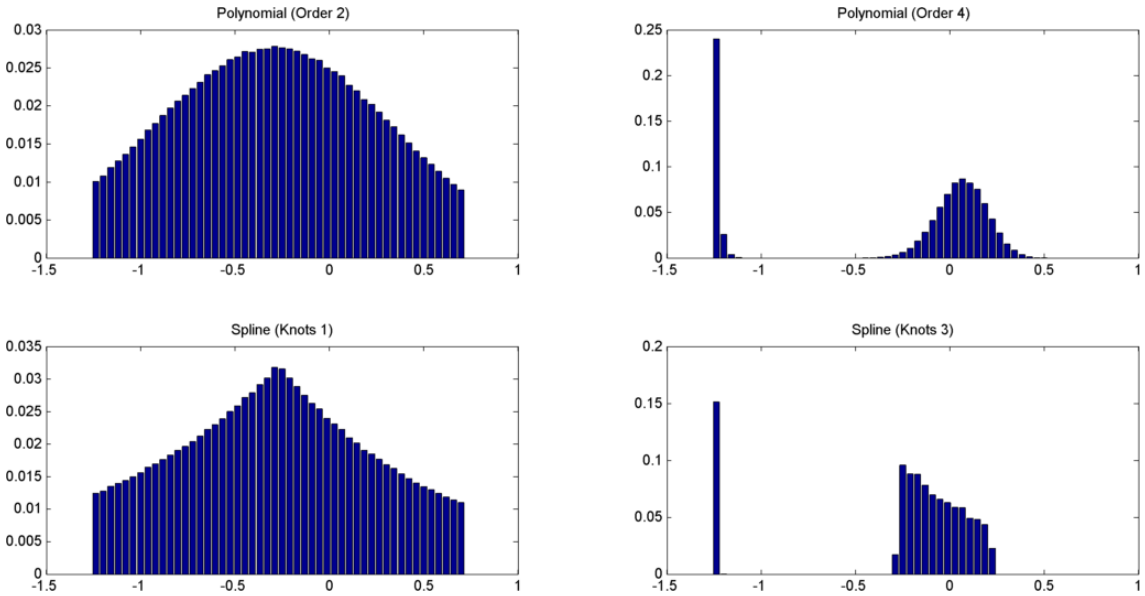

Figure 1.2: Histogram of Marginal Utility of Fuel Cost

Table 1.5 compares computation time of the different LML specifications. In the standard LML model (LML-1), data setup time is higher than that of LML-

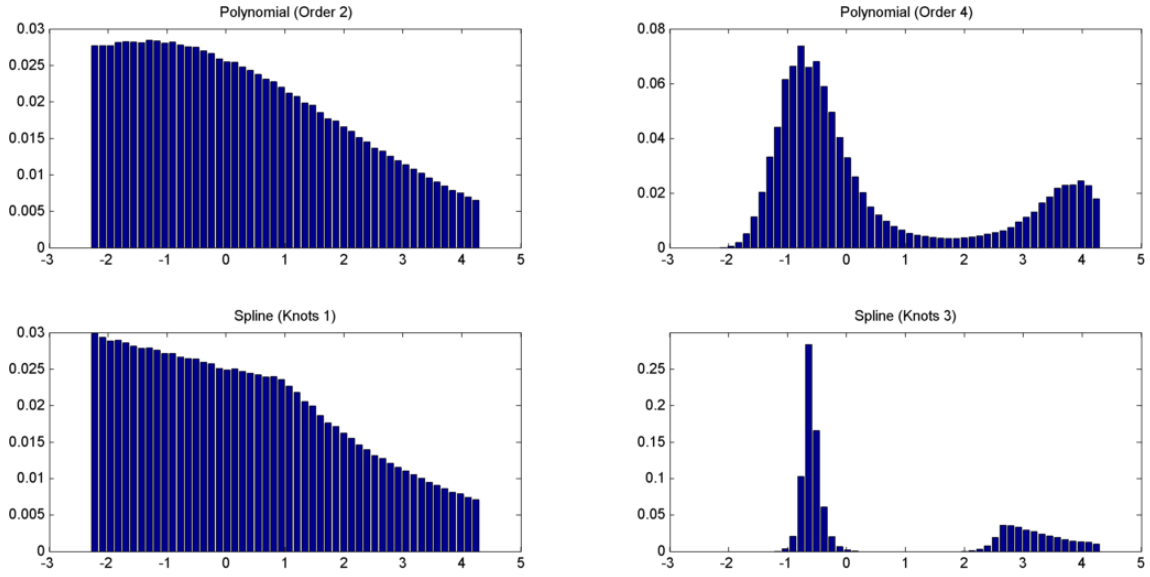

Figure 1.3: Histogram of Marginal Utility of Fuel Availability

FR, because set up in the case of all random parameters also includes the one-time calculation of $L_n(\boldsymbol{\beta}_n)$ (see equation 1.2). This calculation is not needed in the LML-FR case, but we recall that the probability of the sequence of choices needs to be computed at each iteration. This iterative computation of $L_n(\boldsymbol{\beta}_n)$ in LML-FR-4 elevates the estimation time per iteration to 31.2 and 55.7 times (and total estimation time to 16.6 and 41.9 times) that of LML-1 for the polynomial (second order) and spline (knot 1), respectively. These ratios of estimation time remain the same for more flexible specifications. In LML-FR-4, specifying the mixing distribution with a fourth-order polynomial and three-knot spline took around 2.1 and 2.9 times that of second-order polynomial and one-knot spline, respectively.

Table 1.5: Comparing Computation Time of LML and LML-FR

	Data Setup time (Minutes)	Estimation Time (Minutes)	Iterations	Function Evaluations
LML-1 (Poly-order-2)	0.254	1.13	114	144
LML-FR-4 (Poly-order-2)	0.123	22.92	74	76
LML-1 (Poly-order-4)	0.454	3.13	280	335
LML-FR-4 (Poly-order-4)	0.196	48.82	195	205
LML-1 (Spline-knot-1)	0.070	0.41	75	77
LML-FR-4 (Spline-knot-1)	0.028	20.09	66	68
LML-1 (Spline-knot-3)	0.126	1.34	145	151
LML-FR-4 (Spline-knot-3)	0.049	58.81	224	265

1.4 Conclusions

The recently proposed logit-mixed logit (LML) model (Train, 2016) uses a semi-parametric representation of unobserved preference heterogeneity that is simple to implement. This chapter extends LML as derived by Train (2016) –all parameters of interest assumed random– to an LML model with a combination of fixed and random parameters (LML-FR), implemented in preference space, and motivated by the consideration fixed components such as alternative specific constants and interactions between preference parameters and socio-demographics. Whereas computational efficiency is the key feature of LML, we have shown in this chapter that the incorporation of fixed parameters considerably increases estimation time and discussed what provokes the efficiency loss: repeated computation of choice probabilities in iterative optimization. In fact, in the empirical application (stated preferences for alternative fuel vehicles by Chinese consumers; N=578, T = 6, J = 4), estimation time of the LML model with fixed and random

parameters is 20-40 times higher than that of the LML model with all parameters random. This computation time ratio depends on the sample size and can become excruciating for large sample sizes.

The higher LML-FR computation time is also a concern for hypothesis testing and interval estimation, because bootstrapping is the only way to derive LML standard errors[7]. Even if LML-FR loses computational efficiency, its flexibility is invaluable for empirical applications. As illustrated in this study, LML-FR can retrieve multimodality of unobserved preference heterogeneity (c.f. standard parametric models that dominate discrete choice modeling that impose unimodality). The number of parameters and functional form of the mixing distribution, evaluating histogram of random parameters in several scenarios, should be considered for specification selection in addition to standard metrics such as BIC (Bansal et al., 2018*a*).

[7] If 100 bootstrap samples are taken and LML-FR estimation time for one sample is 15 times that of the standard LML estimation time, then computation time of the LML-FR standard errors ends up being 1500 times higher than that of the LML model with all parameters random.

CHAPTER 2

MINORIZATION-MAXIMIZATION (MM) ALGORITHMS FOR SEMIPARAMETRIC LOGIT MODELS: BOTTLENECKS, EXTENSIONS, AND COMPARISONS

2.1 Introduction

2.1.1 Background

With the increase in computation power during the last decade, the mixed multinomial logit (MMNL) model – a random parameter logit model with parametric and continuous heterogeneity distributions – is the most commonly used flexible discrete-choice specification (Train, 2009; McFadden and Train, 2000). The problem of correctly specifying the heterogeneity (or mixing) distribution of the random parameters has received great attention (Hensher and Greene, 2003); however, there is still no consensus among researchers: restricting the shape of the mixing distributions can result into wrong signs and overestimation of welfare measures. Wrong (welfare) estimates can misguide policy and marketing decisions (Fosgerau, 2006; Cherchi and Polak, 2005). To overcome the problem of presuming the shape of the mixing distribution, differing specifications with semi- or nonparametric mixing distributions have been proposed. Vij and Krueger (2017) and Bhat and Lavieri (2017) provide a detailed review of advancements in parametric and semiparametric mixing distributions under extreme-value-distributed

(logit kernel) and normally-distributed (probit kernel) error structures. In general, estimation of these flexible[1] models is complex and computationally expensive.

This study focuses on estimation of two state-of-the-art semiparametric logit models, namely the logit-mixed-logit (LML) and mixture-of-normals multinomial logit (MON-MNL) models, especially in the context of the promising performance of an alternative iterative optimization method with minimal coding, the minorization-maximization algorithm that will be introduced below, as reported for mixed logit (James, 2017).

The logit-mixed-logit model (Train, 2016) generalizes many previous semiparametric models including Bajari et al. (2007), Fosgerau and Bierlaire (2007), Train (2008), and Fox et al. (2011) (cf. Bhat, 1997). In LML, a finite parameter space is divided into a discrete multidimensional grid (cf. Train, 2008). Whereas Train (2008) considers the probability mass at each discrete point as a parameter of interest, LML reduces the number of parameters by specifying this probability using a logit link. In Monte Carlo studies, Bansal et al. (2018*a*) and Franceschinis et al. (2017) successfully tested flexibility of LML as the model could retrieve a series of continuous parametric mixing distributions (bi-modal, tri-modal, log-normal, and uniform) much better than parametric counterparts. The maximum simulated likelihood estimator (MSLE) of LML is much faster than that of parametric models, but computation of standard errors requires bootstrapping. Furthermore, the computational efficiency of point estimation is lost by a factor of 15

[1]Model flexibility understood as the capacity to represent unobserved preference heterogeneity.

to 30 when fixed parameters are introduced (Bansal et al., 2018*b*), and computational efficiency becomes much worse when standard errors are derived.

The mixture-of-normals multinomial logit also offers a flexible representation of unobserved preference heterogeneity. The premise of MON-MNL[2] is that any continuous distribution can be approximated to a given degree of accuracy by a discrete mixture of normals (Ferguson, 1973). Prespecifying the number of mixture components (or classes[3]) imposes a heterogeneity structure, but unlike LML there is no need of predefining the parameter space. Resource-intensive bootstrapping to compute standard errors is not needed in MON-MNL either. In a Monte Carlo study, Fosgerau and Hess (2009) found that MON-MNL outperformed parametric specifications in all scenarios, ranging from retrieving the most trivial uniform distribution to the most complex multimodal distribution. Keane and Wasi (2013) further supported the superiority of MON-MNL in an extensive study of 10 stated preference datasets. However, only a handful of empirical studies have used MON-MNL, possibly due to the complexity of the analytical gradient of the loglikelihood and covergence problems when using numerical gradients. For instance, Fosgerau and Hess (2009) pointed out that MSLE led into troubles for more than 2 normal components in the mixing distribution. Whereas Keane and Wasi (2013) did not explicitly mention any such estimation problem, the authors set bounds on some parameters and also imposed hard constraints on the variance-covariance matrix of each component of the mixture.

[2]MON-MNL was labeled *Mixed-Mixed Logit* by Keane and Wasi (2013) and Latent Class Mixed Multinomial Logit model by Greene and Hensher (2013).

[3]The mixture components can also be interpreted as classes as in a latent class logit model.

Among frequentist methods to estimate logit models, researches have explored iterative optimation methods. Within this class of methods, the expectation-maximization (EM) algorithm has been reported (Bhat, 1997; Cherchi and Guevara, 2012; Sohn, 2017) to outperform MSLE in numerical stability (i.e., less sensitivity to initial values), empirical identification (i.e., avoiding an invertible Hessian matrix), and estimation simplicity. Whereas MSLE directly maximizes the loglikelihood function using quasi-Newton methods, the simplicity of EM stems from iteratively maximizing a simpler surrogate function and update parameters while maintaining monotonic improvements in the loglikelihood (Dempster et al., 1977; McLachlan and Krishnan, 2007). Furthermore, iterative parameter updates of the EM algorithm are either closed-form or straightforward econometric problems that can be solved using standard statistical packages (Train, 2008; Sohn, 2017). EM also provides a convenient parameterization of the complete-data likelihood function without worrying about over-identification (Ruud, 1991). In addition to these nice statistical properties, EM also converges quickly to the neighborhood of the optima. However, EM is plagued by slower convergence within the optimum neighborhood (Dempster et al., 1977). In fact, the computational performance of EM largely hinges upon the underlying data generating process and how well EM re-characterizes the objective function. More specifically, if the complete data model provides much more information about the parameter than the incomplete data model, then the EM algorithm is generally slow (Meilijson, 1989). Ruud (1991) suggested designing hybrid algorithms such that EM starts the maximization process and a Newton-type algorithm finishes it. In fact, Bhat (1997)

could achieve computational efficiency and numerical stability in latent class logit estimation by shifting from EM to quasi-Newton methods when the difference in the loglikelihood of successive iterations achieved a given precision.

For some model specifications EM does not provide closed-form updates (the source of the EM benefits) for all parameters, making EM a rather slow method for estimation. For this reason, researchers have been exploring other alternative estimation methods. EM is actually nested in the minorization-maximization (MM) family of iterative optimization methods (Lange et al., 2000). Note that if all iterative parameter updates in EM are closed-form (optimization-free), MM and EM are basically the same method. MM as proposed by James (2017) replaces iterative optimization of a weighted MNL model with a closed-form parameter update for a mixed logit specification with fixed and random parameters. The MM algorithm in that context was 5-8 times faster than standard EM in general, and outperformed MSLE in some panel-data settings (James, 2017).

2.1.2 Research Gap and Contribution

Whereas Train (2008) proposed EM for an MON-MNL specification with all marginal-utility parameters being random, no such algorithm exists for LML. Even if EM for the all-random-parameters MON-MNL requires just computing sums and multiplications to iteratively update parameters, it has not been further used in practice. Moreover, the consideration of some parameters being fixed may

play against simplicity of estimation of flexible discrete choice models, as noted for EM in the mixed logit work of James (2017).

In addition to focus on a general utility specification with fixed and random parameters due to the implementation challenges that may appear in iterative optimization methods, we argue that inclusion of fixed parameters is important for empirical reasons. Ruud (1996) suggests to hold at least one coefficient fixed in the mixed logit model because a specification with all random coefficients is almost unidentified empirically. Moreover, Train (2009) recommends to keep alternative-specific constants fixed due to the same reason. Empirical identification also restricts parameters to be fixed in the case of interactions with sociodemographics that represent taste variation with respect to a (random) parameter.[4] Finally, willingness to pay for a specific covariate may not vary across the population and thus the estimated parameter may be just non-random.

The *main contribution* of this chapter is to derive MM algorithms (including standard EM) to estimate semiparametric logit models under general utility specifications, identify key bottlenecks in the MM algorithm with all closed-form updates, propose a faster-MM algorithm and illustrate its parallel implementation, and finally compare the different variants of EM and MM algorithms with quasi-Newton methods in a Monte Carlo study and an empirical study. Since our contribution is fourfold, details are discussed below.

[4]Consider an indicator variable for males D_{male} and a taste-variation-specification of the type $(\beta_{ik} + \beta_{\text{male}} D_{\text{male}}) x_{ik}$. While β_{ik} can be considered random, for empirical identification β_{male} would be considered fixed in applied work.

The *first contribution* of this study is to derive standard EM for LML (cf. Bhat, 1997, for a latent class logit model) and to extend EM for MON-MNL under more general utility specifications. Even though the proposed EM algorithm for LML and MON-MNL can be implemented easily in any standard statistical package, we show that – just as in mixed logit, although the equations are different – EM intrinsically involves computationally burdensome optimizations of weighted MNL (WMNL) loglikelihood functions (see sections 2.2.2 and 2.3.2). Whereas the EM algorithm for LML involves iterative estimation of two WMNL models, we show that EM for MON-MNL requires as many WMNL loglikelihood maximizations at each iteration as the number of mixture components. We also show that these resource-intensive computations may even rule out EM as a feasible option to estimate semiparametric logit models in practice.

Given the computational performance issues of EM discussed above, the *second contribution* of this study is to derive MM for both LML and MON-MNL models. We discuss and illustrate several advantages of MM over EM in section 2.4. In particular, optimization-free (closed-form) parameter updates in MM makes it attractive since it can be easily integrated into flexible[5] estimation software. Furthermore MM just requires storage of sufficient statistics, and we thus demonstrate how parallel computation can result into 80% reduction in estimation time.[6] Although we were first expecting to see computational efficiency improvements

[5]Code flexibility is understood as the capacity of software to allow the analyst to directly specify the desired loglikelihood, avoiding restrictions such as only linear-in-parameter utility specifications.

[6]The EM algorithm can also be parallelized, but storage and communication of large multidimensional arrays neutralizes the computational gains of parallelization.

such as the ones observed by James (2017) for mixed logit, MM computation time for LML was disappointing even after parallel estimation. What happens is that the MM algorithm's lower bound approximation is very poor if the WMNL model has a large choice set (see Böhning and Lindsay, 1988), which is by construction exactly the case of LML: the number of "alternatives" in the WMNL update is equal to the number of random draws (in the order of 1000s) taken from the multidimensional grid that represents the parameter space. In fact, through a Monte Carlo study, we found that MM for MON-MNL and even the parametric mixed logit model also encounter computational failure for a large choice set. This issue is critical because large choice sets are often encountered in revealed preference studies, precisely in applied work where alternative estimation methods such as EM and MM are needed (von Haefen and Domanski, 2018).

The *third contribution* of this study is the identification of bottlenecks in the implementation of MM and provide improvements. Following the method suggested by Böhning and Lindsay (1988), we propose a *faster-MM* algorithm in which the embedded step-size is corrected. The proposed faster-MM algorithm is general and useful to improve the lower-bound approximation of any WMNL loglikelihood while keeping the simplicity of MM. However, the computational gains largely depend on the cardinality of the choice set. For example, faster-MM could reduce computation time of MM for an LML model by a factor of around 100 due to a very high (order of 1000s) cardinality of the choice set of the iterative WMNL.

The *fourth contribution* of this study is to highlight shortcomings and advantages of the existing and proposed estimation methods using an extensive Monte Carlo study and an empirical application to estimate consumer's willingness to adopt electric motorcycles in Solo, Indonesia.

The remaining of the chapter is organized as follows: sections 2 and 3 derive the iterative-optimization estimation algorithms for LML and MON-MNL; section 4 draws insights from the derived procedures; section 5 describes the Monte Carlo studies and discusses the findings; section 6 focuses on the empirical study; and conclusions are detailed in section 7.

2.2 Iterative Optimization Methods to Estimate the Logit Mixed Logit (LML) Model

2.2.1 Logit-Mixed Logit (LML)

Let N be the number of decision-makers in a sample where each agent faces T choice situations and chooses a utility-maximizing alternative from a set of J alternatives. The conditional indirect utility of decision-maker i from making choice j in choice situation t is:

$$U_{itj} = \boldsymbol{x}_{itj}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i + \varepsilon_{itj}, \tag{2.1}$$

where $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, J\}$, and $t \in \{1, \ldots, T\}$. We consider the general case in which attributes can be partitioned between those having fixed parameters ($\boldsymbol{x}$) and those with random parameters ($\boldsymbol{z}$) with a general continuous heterogeneity distribution.[7] The attribute vector $\boldsymbol{x}_{itj}$ thus has a fixed preference parameter vector $\boldsymbol{\alpha}$, whereas $\boldsymbol{z}_{itj}$ has a random, individual-specific preference vector $\boldsymbol{\beta}_i$. The preference shock ε_{itj} is independent across individuals, choices and time, and is identically distributed Type-I Extreme Value. Thus, the probability of choosing alternative j by individual i in choice situation t, conditional on $\boldsymbol{\beta}_i$, has a logit link:

$$P_{itj}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \frac{\exp\left(\boldsymbol{x}_{itj}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_{itk}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itk}^T \boldsymbol{\beta}_i\right)}. \tag{2.2}$$

If individual i chooses alternative j in choice situation t, one can define the choice indicator $d_{itj} = \mathbb{I}(\text{j chosen}|i, t)$. For the sequence of choices made by individual i, the conditional likelihood $\mathcal{L}_i(\boldsymbol{\alpha}|\boldsymbol{\beta}_i)$ is:

$$\mathcal{L}_i(\boldsymbol{\alpha}|\boldsymbol{\beta}_i) = \prod_{t=1}^{T} \prod_{j=1}^{J} [P_{itj}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i)]^{d_{itj}}. \tag{2.3}$$

In the LML model, preferences variations are represented using a discrete mixing distribution over a finite support set S. The probability of the random $\boldsymbol{\beta}_i$ being equal to a specific value $\boldsymbol{\beta}_{ir}$ is represented by the following logit link (hence the logit-mixed logit name proposed by Train, 2016):

$$w_i(\boldsymbol{\beta}_i = \boldsymbol{\beta}_{ir}|\boldsymbol{\phi}) = \frac{\exp\left(\boldsymbol{y}(\boldsymbol{\beta}_{ir})^T \boldsymbol{\phi}\right)}{\sum_{s \in S} \exp\left(\boldsymbol{y}(\boldsymbol{\beta}_{is})^T \boldsymbol{\phi}\right)}, \tag{2.4}$$

[7] As discussed in the introduction, the consideration of a combination of fixed and random parameters not only is empirically justified but also offers estimation challenges that need to be addressed.

where $\boldsymbol{\phi}$ is a vector of parameters and $\boldsymbol{y}(\boldsymbol{\beta}_r)$ is a vector-valued function (e.g. spline, step, or polynomial function) that captures the shape of the mixing distribution.

If $\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \boldsymbol{\phi}\}$ summarizes the parameters of interest, the unconditional likelihood $P_i(\boldsymbol{\psi})$ of agent i is:

$$P_i(\boldsymbol{\psi}) = \sum_{r \in S} \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir} | \boldsymbol{\phi}), \tag{2.5}$$

and the corresponding loglilkelihood $\ell(\boldsymbol{\psi})$ of the sample is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{N} \ln \Big(\sum_{r \in S} \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir} | \boldsymbol{\phi}) \Big). \tag{2.6}$$

Maximizing the loglikelihood becomes intractable in practice if the entire support set S is used. Therefore, a large subset of parameter vectors (for example, 2000 vectors) for each decision-maker is sampled from the support set (see sections 2.5.1 and 2.6 for details) to derive the maximum simulated likelihood estimator of the model. The standard errors of the parameters of interest are calculated through bootstrapping.

2.2.2 LML Estimation using the EM Algorithm

The EM algorithm was originally developed to deal with missing data (Dempster et al., 1977). Since several loglikelihood maximization problems can be viewed as a missing data problem, the EM algorithm has been widely used in different disciplines (McLachlan and Krishnan, 2007). The EM algorithm is a two-step –

the *expectation* (E) step and the *maximization* (M) step – iterative procedure. In the E-step, data are completed (in a probabilistic sense) conditional on previous iteration parameters. In the M-step, parameters are optimized conditional on the completed data. The algorithm terminates when difference in parameter estimates of two consecutive iterations is very small.

Bhat (1997) first introduced the EM algorithm into the discrete choice literature for the estimation of models with endogenous segmentation or latent classes. Since LML generalizes the latent class logit model, the missing data for the EM algorithm in LML estimation are the parameters of each agent in each draw ($\boldsymbol{\beta}_{ir}$) just as originally suggested in Bhat (1997). As a result, the M-step $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^m)$ and E-step of the EM algorithm for LML model are:

$$\text{E-step}: \quad h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) = \frac{\mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi}^m)}{\sum_{r\in S} \mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi}^m)} = \frac{\mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi}^m)}{P_i(\boldsymbol{\psi}^m)}, \tag{2.7}$$

$$\text{M-step}: \quad Q(\boldsymbol{\psi}|\boldsymbol{\psi}^m) = \sum_{i=1}^{N} \sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) \ln\left(\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi})\right), \tag{2.8}$$

where $h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)$ are weights that are computed at each iteration ($\boldsymbol{\psi}_{m+1}$) using the previous iteration ($\boldsymbol{\psi}_m$). The M-step surrogate function is additive separable in the fixed parameters ($\boldsymbol{\alpha}$) and the parameters defining the shape of the mixing distribution ($\boldsymbol{\phi}$):

$$Q(\boldsymbol{\psi}|\boldsymbol{\psi}^m) = \sum_{i=1}^{N} \sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) \ln(\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir})) + \sum_{i=1}^{N} \sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) \ln(w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi})). \tag{2.9}$$

As expected, optimizing these two independent functions separately is much easier than the joint maximization problem. The EM update equations are:

$$Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^m) = \sum_{i=1}^{N} \sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) ln(\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir})) \implies \boldsymbol{\alpha}^{m+1} = \underset{\boldsymbol{\alpha}}{\text{argmax}}\ Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^m) \tag{2.10}$$

$$Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m) = \sum_{i=1}^{N} \sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) ln(w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi})) \implies \boldsymbol{\phi}^{m+1} = \underset{\boldsymbol{\phi}}{\text{argmax}}\ Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m), \tag{2.11}$$

where $Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^m)$ represents a weighted multinomial logit loglikelihood (with fixed utility weights $\boldsymbol{\alpha}$) and $Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m)$ represents a very similar expression to the weighted multinomial logit loglikelihood (with characteristics $\boldsymbol{y}(\boldsymbol{\beta}_{ir})$ and weights $\boldsymbol{\phi}$). If all utility parameters are random, the algorithm remains the same after removing equation 2.10 to update the fixed parameters. The complete EM algorithm for the LML model is summarized below:

2.2.3 LML Estimation using the MM Algorithm

In the EM algorithm for LML, numerical maximization of $Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^m)$ and $Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m)$ at each iteration can be computationally burdensome. Inspired in the work by James (2017) for mixed logit, we propose the use of the minorization-maximization (MM) algorithm (Lange et al., 2000), where closed-form surrogate functions $[\widetilde{Q}(\boldsymbol{\alpha}|\boldsymbol{\psi}^m), \widetilde{Q}(\boldsymbol{\phi}|\boldsymbol{\psi}^m)]$ create closed-form updates of the parameters ($\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \boldsymbol{\phi}\}$) without solving any optimization problem.

Algorithm 1: EM for the LML Model

Initialization
For each i, draw $\boldsymbol{\beta}_{ir}$, $r = 1, \ldots, R$ (e.g., $R = 2000$), from the support set S;
Compute $\boldsymbol{y}(\boldsymbol{\beta}_{ir})$ using sieve functions such as spline;
Initialize parameters $m = 0 : \boldsymbol{\psi}^0 = \{\boldsymbol{\alpha}^0, \boldsymbol{\phi}^0\}$
while $\|\boldsymbol{\psi}^{m+1} - \boldsymbol{\psi}^m\|_\infty < Tol.$ **do**
  Step 1: Calculation of the weight $[h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)]$;
  Calculate $P_{itj}(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir})$ for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.2;
  Calculate $\mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir})$ for each i and for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.3;
  Calculate $w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi}^m)$ for each i and for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.4;
  Calculate $h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)$ for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.7;
  Step 2: Update parameters;
  Update $\boldsymbol{\alpha}^{m+1}$ using Eq. 2.10;
  Update $\boldsymbol{\phi}^{m+1}$ using Eq. 2.11;
end

The new surrogate functions are derived using a quadratic lower bound approximation of the Hessian.[8] This approximation can be used to reduce the optimization burden of any EM algorithm which iteratively optimizes the loglikelihood of a weighted MNL model.[9]

Updating $\boldsymbol{\phi}$

The new surrogate function to update $\boldsymbol{\phi}$ using the approximation is:

$$\widetilde{Q}(\boldsymbol{\phi}|\boldsymbol{\psi}^m) = Q(\boldsymbol{\phi}^m|\boldsymbol{\psi}^m) + (\boldsymbol{\phi} - \boldsymbol{\phi}^m)^T \boldsymbol{g}_{\boldsymbol{\phi}}^m + \frac{(\boldsymbol{\phi} - \boldsymbol{\phi}^m)^T \boldsymbol{B}_{\boldsymbol{\phi}}^m (\boldsymbol{\phi} - \boldsymbol{\phi}^m)}{2}, \tag{2.12}$$

[8] A function can be bounded below by a quadratic approximation if there exists a global lower bound to the second derivative (Böhning and Lindsay, 1988). In fact, James (2017) proposed the MM algorithm to estimate the MMNL model exploiting the same idea.

[9] However, MM estimation time can be higher than EM estimation time, or vice versa, depending on the tightness of the Hessian approximation.

$$\text{where} \boldsymbol{g}_{\boldsymbol{\phi}}^{m} = \left.\frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\psi}^{m})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{m}} \quad \text{and} \quad \boldsymbol{B}_{\boldsymbol{\phi}}^{m} \leq \frac{\partial^{2} Q(\boldsymbol{\phi}|\boldsymbol{\psi}^{m})}{\partial \boldsymbol{\phi}^{2}}.$$

Thus,

$$\boldsymbol{\phi}^{m+1} = \boldsymbol{\phi}^{m} - [\boldsymbol{B}_{\boldsymbol{\phi}}^{m}]^{-1} \boldsymbol{g}_{\boldsymbol{\phi}}^{m}. \tag{2.13}$$

This new update equation of $\boldsymbol{\phi}$ is a single Newton step that only requires the gradient and lower bound on the Hessian of the EM surrogate function $Q(\boldsymbol{\phi}|\boldsymbol{\psi}^{m})$, which are computed as follows (see Böhning and Lindsay, 1988, for details):

$$\boldsymbol{g}_{\boldsymbol{\phi}}^{m} = \left.\frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\psi}^{m})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{m}} = \sum_{i=1}^{N} \sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m})(\boldsymbol{y}(\boldsymbol{\beta}_{ir}) - \sum_{v \in S} [\boldsymbol{y}(\boldsymbol{\beta}_{iv}) w_{i}(\boldsymbol{\beta}_{iv}|\boldsymbol{\phi}^{m})]) \tag{2.14}$$

$$\begin{aligned} \boldsymbol{H}_{\boldsymbol{\phi}}^{m} = \left.\frac{\partial^{2} Q(\boldsymbol{\phi}|\boldsymbol{\psi}^{m})}{\partial \boldsymbol{\phi}^{2}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{m}} = -\sum_{i=1}^{N} \sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m}) \Big[\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv})[\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} w_{i}(\boldsymbol{\beta}_{iv}|\boldsymbol{\phi}^{m}) - \\ \Big(\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv}) w_{i}(\boldsymbol{\beta}_{iv}|\boldsymbol{\phi}^{m}) \Big) \Big(\sum_{v \in S} [\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} w_{i}(\boldsymbol{\beta}_{iv}|\boldsymbol{\phi}^{m}) \Big) \Big] \end{aligned} \tag{2.15}$$

$$\boldsymbol{B}_{\boldsymbol{\phi}}^{m} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m}) \Big[\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv})[\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} - \frac{1}{R} \Big(\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv}) \Big) \Big(\sum_{v \in S} [\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} \Big) \Big] \tag{2.16}$$

$$\sum_{r \in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m}) = 1 \implies \tag{2.17}$$

$$\boldsymbol{B}_{\boldsymbol{\phi}}^{m} = \boldsymbol{B}_{\boldsymbol{\phi}} = -\frac{1}{2} \sum_{i=1}^{N} \Big[\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv})[\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} - \frac{1}{R} \Big(\sum_{v \in S} \boldsymbol{y}(\boldsymbol{\beta}_{iv}) \Big) \Big(\sum_{v \in S} [\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T} \Big) \Big].$$

Updating $\boldsymbol{\alpha}$

Similarly we take lower bound approximation of $Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^{m})$ (see equation 2.12) and the update equation for $\boldsymbol{\alpha}$ is:

$$\boldsymbol{\alpha}^{m+1} = \boldsymbol{\alpha}^{m} - [\boldsymbol{B}_{\boldsymbol{\alpha}}^{m}]^{-1} \boldsymbol{g}_{\boldsymbol{\alpha}}^{m}. \tag{2.18}$$

The gradient $\boldsymbol{g}_{\alpha}^{m}$ in this case is:

$$\boldsymbol{g}_{\alpha}^{m} = \left.\frac{\partial Q(\boldsymbol{\alpha}|\boldsymbol{\psi}^{m})}{\partial \boldsymbol{\alpha}}\right|_{\alpha=\alpha^{m}} = \sum_{i=1}^{N}\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m})\Big[\sum_{t=1}^{T}\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big(d_{itj} - P_{itj}(\boldsymbol{\alpha}^{m},\boldsymbol{\beta}_{ir})\Big)\Big], \tag{2.19}$$

and the Hessian:

$$\boldsymbol{H}_{\alpha}^{m} = -\sum_{i=1}^{N}\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m})\Big\{\sum_{t=1}^{T}\Big[\sum_{j=1}^{J}\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}{}^{T}P_{itj}(\boldsymbol{\alpha}^{m},\boldsymbol{\beta}_{ir}) - \Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}P_{itj}(\boldsymbol{\alpha}^{m},\boldsymbol{\beta}_{ir})\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}P_{itj}(\boldsymbol{\alpha}^{m},\boldsymbol{\beta}_{ir})\Big)^{T}\Big]\Big\}. \tag{2.20}$$

$$\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^{m}) = 1 \implies$$

$$\boldsymbol{B}_{\alpha}^{m} = \boldsymbol{B}_{\alpha} = -\frac{1}{2}\sum_{i=1}^{N}\Big\{\sum_{t=1}^{T}\Big[\sum_{j=1}^{J}(\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}^{T}) - \frac{1}{J}\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big)^{T}\Big]\Big\}. \tag{2.21}$$

All the steps of the MM algorithm to estimate the LML model are given below:

2.2.4 LML Estimation using the Faster-MM Algorithm

It is important to note that the number of 'alternatives' in equation 2.11 (updating $\boldsymbol{\phi}$) that stems from a logit link is equal to the number of draws R (size of the estimation subset) for each agent from the support set S. For a good coverage of the parameter space, R should be large (for example, $R = 2000$). Böhning and Lindsay (1988) highlight the fact that if the number of alternatives is large, the approximation of the Hessian in equation 2.17 becomes $\boldsymbol{B}_{\boldsymbol{\phi}}^{m} = \boldsymbol{B}_{\boldsymbol{\phi}} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{v\in S}\boldsymbol{y}(\boldsymbol{\beta}_{iv})[\boldsymbol{y}(\boldsymbol{\beta}_{iv})]^{T}$,

Algorithm 2: MM for the LML Model

Initialization
For each i, draw $\boldsymbol{\beta}_{ir}$, $r = 1, \ldots, R$ (e.g., $R = 2000$), from the support set S;
Compute $\boldsymbol{y}(\boldsymbol{\beta}_{ir})$ using Sieve functions such as spline;
Compute $\boldsymbol{B}_{\boldsymbol{\phi}}^{-1}$ using Eq. 2.17
Compute $\boldsymbol{B}_{\alpha}^{-1}$ using Eq. 2.21
Initialize parameters $m = 0 : \boldsymbol{\psi}^0 = \{\boldsymbol{\alpha}^0, \boldsymbol{\phi}^0\}$
while $\|\boldsymbol{\psi}^{m+1} - \boldsymbol{\psi}^m\|_\infty < Tol.$ **do**
  Step 1: Calculation of the weight $[h_{ir}(\boldsymbol{\beta}_{ir}|\psi^m)]$;
  Calculate $P_{itj}(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir})$ for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.2;
  Calculate $\mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{ir})$ for each i and for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.3;
  Calculate $w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\phi}^m)$ for each i and for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.4;
  Calculate $h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)$ for each $\boldsymbol{\beta}_{ir}$ using Eq. 2.7 ;
  Step 2: Update parameters ;
  Compute $\boldsymbol{g}_{\boldsymbol{\phi}}^m$ using Eq. 2.14 and update $\boldsymbol{\phi}^{m+1}$ using Eq. 2.13;
  Compute $\boldsymbol{g}_{\alpha}^m$ using Eq. 2.19 and update $\boldsymbol{\alpha}^{m+1}$ using Eq. 2.18;
end

which is a very crude approximation. This observation is illustrated in the Appendix B using a sketch of the proof for the lower bound of hessian and also in a Monte Carlo study. Specifically, Böhning and Lindsay (1988) mentions the problem of the curvature of the loglikelihood varying sharply as a function of initial values and direction. With a modified step-size, as suggested by the authors, for a broad family of statistical models, we propose the following simple algorithmic improvement to update $\boldsymbol{\phi}$:

Step 1: Compute the step size for MM: $\boldsymbol{\mu}_{\boldsymbol{\phi}}^m = -[\boldsymbol{B}_{\boldsymbol{\phi}}^m]^{-1}\boldsymbol{g}_{\boldsymbol{\phi}}^m$.

Step 2: Modify the step size: $\boldsymbol{\zeta}_{\boldsymbol{\phi}}^m = \eta_{\boldsymbol{\phi}}^m \boldsymbol{\mu}_{\boldsymbol{\phi}}^m$.

Step 3: Update $\boldsymbol{\phi}$: $\boldsymbol{\phi}^{m+1} = \boldsymbol{\phi}^m + \boldsymbol{\zeta}_{\boldsymbol{\phi}}^m$.

Intuitively, in this faster-MM algorithm the original step size $\boldsymbol{\mu}_{\boldsymbol{\phi}}^m$ is augmented by a positive multiplier $\eta_{\boldsymbol{\phi}}^m$, and the modified step size $\boldsymbol{\zeta}_{\boldsymbol{\phi}}^m$ is then used to update $\boldsymbol{\phi}$. The use of $\boldsymbol{\zeta}_{\boldsymbol{\phi}}$, instead of $\boldsymbol{\mu}_{\boldsymbol{\phi}}^m$, not only maintains monotonic improvements in the loglikelihood, but also ensures fast convergence of the MM algorithm for LML (see simulation results for LML in section 2.5.1). This faster-MM method can be extended to improve the convergence rate of MM estimation of logit-type models with large choice sets in general:[10] the MM algorithm for mixed logit as originally implemented by James (2017) is actually extremely slow if the number of alternatives is large and our proposed faster-MM algorithm can provide significant improvements (see section 2.5.2 for the simulation results). We also derive the faster-MM for MON-MNL (section 2.3.4).

Computation of $\eta_{\boldsymbol{\phi}}^m$

Böhning and Lindsay (1988) suggest writing a standard Taylor series expansion and then solving for the scalar multiplier. We derive the expression for $\eta_{\boldsymbol{\phi}}^m$ exactly following those steps:

$$Q(\boldsymbol{\phi}^{m+1}) - Q(\boldsymbol{\phi}^m) = Q(\boldsymbol{\phi}^m + \eta_{\boldsymbol{\phi}}^m \boldsymbol{\mu}_{\boldsymbol{\phi}}^m) - Q(\boldsymbol{\phi}^m) \geq \eta_{\boldsymbol{\phi}}^m (\boldsymbol{\mu}_{\boldsymbol{\phi}}^m)^T \boldsymbol{g}_{\boldsymbol{\phi}}^m + \frac{(\eta_{\boldsymbol{\phi}}^m)^2 (LB)}{2}. \tag{2.22}$$

Solving for $\eta_{\boldsymbol{\phi}}^m$:

$$\eta_{\boldsymbol{\phi}}^m = -\frac{(\boldsymbol{\mu}_{\boldsymbol{\phi}}^m)^T \boldsymbol{g}_{\boldsymbol{\phi}}^m}{LB}, \tag{2.23}$$

[10] For example, in the case of 4 alternatives, the multiplier $\eta_{\boldsymbol{\phi}}^m$ can be of order 1.2, which is insignifcant for improving computational efficiency.

where LB is a lower bound on the quadratic form of the Hessian that can be calculated as:

$$LB = -\sum_{i=1}^{N}\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)LB^i_{\boldsymbol{\phi}},$$

where $LB^i_{\boldsymbol{\phi}}$ is a lower bound on the quadratic form of $\boldsymbol{H}^m_{\boldsymbol{\phi},ir}$ and is calculated as follows[11]:

$$\begin{aligned} &LB^i_{\boldsymbol{\phi}} = .5(m_i^2 + M_i^2 - .5((m_i + M_i)^2)) \\ &\text{where} \quad m_i = \min_{1\le v\le R}((\boldsymbol{\mu}^m_{\boldsymbol{\phi}})^T y(\boldsymbol{\beta}_{iv})) \quad \text{and} \quad M_i = \max_{1\le v\le R}((\boldsymbol{\mu}^m_{\boldsymbol{\phi}})^T y(\boldsymbol{\beta}_{iv})). \end{aligned} \tag{2.24}$$

Recalling that the Hessian is:

$$\boldsymbol{H}^m_{\boldsymbol{\phi}} = \left.\frac{\partial^2 Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m)}{\partial\boldsymbol{\phi}^2}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^m} = -\sum_{i=1}^{N}\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)\boldsymbol{H}^m_{\boldsymbol{\phi},ir}, \tag{2.25}$$

and recognizing that $\sum_{r\in S} h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m) = 1$, it is possible to implement the lower bound as:

$$LB = -\sum_{i=1}^{N} LB^i_{\boldsymbol{\phi}}. \tag{2.26}$$

Note that Böhning and Lindsay (1988) suggested to use $LB = \sum_{i=1}^{N} LB^i_{\boldsymbol{\phi}}$ in the original paper for specifications such as the Cox proportional hazards model. When we first implemented the bound without the negative sign, the MM algorithm lost monotonicity and was not converging. In fact, the loglikelihood was fluctuating randomly, instead of increasing at each iteration. We soon realized that monotonocity would be ensured if $LB = -\sum_{i=1}^{N} LB^i_{\boldsymbol{\phi}}$. With the corrected sign of LB, as shown in equation 2.26, the MM algorithm not only converged but also

[11] See equation 2.15 and equations 5.5 - 5.7 of Böhning and Lindsay (1988) for further information.

convergence was achieved much faster, as we were expecting. To make sure intuition about the sign of the bound was correct, we provide a formal proof below.

Checking the sign of LB. As mentioned above, irrespective of the chosen sample, the sign of $\eta_{\boldsymbol{\phi}}^m$ must be positive to ensure monotonicity of the algorithm. We now show that $LB = -\sum_{i=1}^{N} LB_{\boldsymbol{\phi}}^i$ (as opposed to the inverse $LB = \sum_{i=1}^{N} LB_{\boldsymbol{\phi}}^i$) fulfills this requirement. In effect,

$$\eta_{\boldsymbol{\phi}}^m = -\frac{(\boldsymbol{\mu}_{\boldsymbol{\phi}}^m)^T \boldsymbol{g}_{\boldsymbol{\phi}}^m}{LB} = -\frac{(-[\boldsymbol{B}_{\boldsymbol{\phi}}^m]^{-1} \boldsymbol{g}_{\boldsymbol{\phi}}^m)^T \boldsymbol{g}_{\boldsymbol{\phi}}^m}{LB} = \frac{(\boldsymbol{g}_{\boldsymbol{\phi}}^m)^T [\boldsymbol{B}_{\boldsymbol{\phi}}^m]^{-1} \boldsymbol{g}_{\boldsymbol{\phi}}^m}{LB}. \tag{2.27}$$

Consider first the numerator $(\boldsymbol{g}_{\boldsymbol{\phi}}^m)^T [\boldsymbol{B}_{\boldsymbol{\phi}}^m]^{-1} \boldsymbol{g}_{\boldsymbol{\phi}}^m \leq 0$: since the objective function is concave, the Hessian is negative semi-definite and thus $[\boldsymbol{B}_{\boldsymbol{\phi}}^m]^{-1}$ is negative semi-definite.

Consider now the denominator LB. Since the objective function is concave, $\frac{\partial^2 Q(\boldsymbol{\phi}|\boldsymbol{\psi}^m)}{\partial \boldsymbol{\phi}^2}$ (see Equation 2.25) is negative semi-definite. Additionally, $h_{ir}(\boldsymbol{\beta}_{ir}|\boldsymbol{\psi}^m)$ is a positive weight and thus $\boldsymbol{H}_{\boldsymbol{\phi},ir}^m$ is a positive semi-definite matrix. Since LB_i is a lower bound on the quadratic form of $\boldsymbol{H}_{\boldsymbol{\phi},ir}^m$, it has to be non-negative. If $LB = \sum_i LB_i$, $LB \geq 0 \implies \eta_{\boldsymbol{\phi}}^m \leq 0$, but if $LB = -\sum_i LB_i$, then $LB \leq 0 \implies \eta_{\boldsymbol{\phi}}^m \geq 0$ as needed. (Q.E.D.)

2.3 Iterative Optimization Methods to Estimate MON-MNL

2.3.1 Mixture-of-normals logit (MON-MNL)

If population has C latent classes (i.e., components in the mixture), utility derived for individual i of class c from making choice j in choice situation t is:

$$U_{itj} = \boldsymbol{x}_{itj}^T \boldsymbol{\alpha}_c + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i^c + \varepsilon_{itj}, \tag{2.28}$$

where $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, J\}$, $t \in \{1, \ldots, T\}$, and $c \in \{1, \ldots, C\}$. The alternative-specific characteristics $\boldsymbol{x}_{itj}$ have a fixed utility weight vector $\boldsymbol{\alpha}_c$, and $\boldsymbol{z}_{itj}$ has an individual-specific parameter vector $\boldsymbol{\beta}_i^c$ (specific to class c). The taste shock ε_{itj} is independent and identically distributed Type-I Extreme Value. For the sequence of choices made by individual i, the conditional likelihood $\mathcal{L}_i(\boldsymbol{\alpha}_c, \boldsymbol{\beta}_i^c)$ is:

$$\mathcal{L}_i(\boldsymbol{\alpha}_c, \boldsymbol{\beta}_i^c) = \prod_{t=1}^{T} \prod_{j=1}^{J} [P_{itj}]^{d_{itj}} = \prod_{t=1}^{T} \prod_{j=1}^{J} \left[\frac{exp(\boldsymbol{x}_{itj}^T \boldsymbol{\alpha}_c + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i^c)}{\sum_{k=1}^{J} exp(\boldsymbol{x}_{itk}^T \boldsymbol{\alpha}_c + \boldsymbol{z}_{itk}^T \boldsymbol{\beta}_i^c)} \right]^{d_{itj}}. \tag{2.29}$$

Consider a latent class membership variable $\boldsymbol{S} = (S_1, \ldots, S_N)$ such that : $P(S_i = c) = s_c \quad i \in \{1, \ldots, N\}$, where $0 \leq s_c \leq 1$ and $\sum_1^C s_c = 1$. Conditional on class membership, the random parameter $\boldsymbol{\beta}_i^c$ is normally distributed with mean $\boldsymbol{\gamma}_c$ and variance-covariance matrix $\boldsymbol{\Delta}_c$. Thus the loglilkelihood $\ell(\boldsymbol{\psi})$ of the sample in terms

of the unconditional likelihood $P_i(\boldsymbol{\psi})$ of individual i is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{N} ln\Big(P_i(\boldsymbol{\psi})\Big) = \sum_{i=1}^{N} ln\Bigg(\sum_{c=1}^{C}\Big\{s_c\Big[\int_{\beta}\mathcal{L}_i(\boldsymbol{\alpha}_c,\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{\gamma}_c,\boldsymbol{\Delta}_c)d\boldsymbol{\beta}\Big]\Big\}\Bigg) \tag{2.30}$$
$$\text{where} \quad \boldsymbol{\psi} = \{\boldsymbol{\alpha}_1, s_1, \boldsymbol{\gamma}_1, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\alpha}_C, s_C, \boldsymbol{\gamma}_C, \boldsymbol{\Delta}_C\}$$

2.3.2 MON-MNL Estimation using the EM Algorithm

Following Train (2008), the E-step and the objective function of the M-step $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^m)$ of the EM algorithm for MON-MNL model are:

$$\text{E-step}: \quad h_{ic}(.|\boldsymbol{\psi}^m) = \frac{s_c^m \mathcal{L}_i(\boldsymbol{\alpha}_c^m,\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{\gamma}_c^m,\boldsymbol{\Delta}_c^m)}{P_i(\boldsymbol{\psi}^m)}$$
$$\text{M-step}: \quad \boldsymbol{\psi}^{m+1} = \underset{\boldsymbol{\psi}}{\text{argmax}} \sum_{i=1}^{N}\sum_{c=1}^{C}\Big[\int_{\beta} h_{ic}(.|\boldsymbol{\psi}^m) ln\Big(s_c\mathcal{L}_i(\boldsymbol{\alpha}_c,\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{\gamma}_c,\boldsymbol{\Delta}_c)\Big)d\boldsymbol{\beta}\Big]. \tag{2.31}$$

In the E-step, $h_{ic}(.|\boldsymbol{\psi}^m)$ can be recognized as weight. The derived parameter M-step update equations are:

$$s_c^{m+1} = \frac{\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)}{\sum_{i=1}^{N}\sum_{v=1}^{C}\sum_{r=1}^{R} h_{ivr}(.|\boldsymbol{\psi}^m)} \tag{2.32}$$

$$\begin{aligned}\boldsymbol{\gamma}_c^{m+1} &= \frac{\sum_{i=1}^{N}\sum_{r=1}^{R}\left[h_{icr}(.|\boldsymbol{\psi}^m)\boldsymbol{\beta}_{icr}^m\right]}{\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)},\\ \boldsymbol{\Delta}_c^{m+1} &= \frac{\sum_{i=1}^{N}\sum_{r=1}^{R}\left[h_{icr}(.|\boldsymbol{\psi}^m)\left\{(\boldsymbol{\beta}_{icr}^m - \boldsymbol{\gamma}_c^{m+1})(\boldsymbol{\beta}_{icr}^m - \boldsymbol{\gamma}_c^{m+1})^T\right\}\right]}{\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)},\end{aligned} \tag{2.33}$$

$$\boldsymbol{\alpha}_c^{m+1} = \underset{\boldsymbol{\alpha}_c}{\text{argmax}}\ Q(\boldsymbol{\alpha}_c|\boldsymbol{\psi}^m) = \underset{\boldsymbol{\alpha}_c}{\text{argmax}} \sum_{i=1}^{N}\sum_{r=1}^{R}\Big[h_{icr}(.|\boldsymbol{\psi}^m) ln\Big(\mathcal{L}_i(\boldsymbol{\alpha}_c,\boldsymbol{\beta}_{icr})\Big)\Big]. \tag{2.34}$$

2.3.3 MON-MNL Estimation using the MM Algorithm

Updating the equation of $\boldsymbol{\alpha}_c$ in the EM algorithm involves optimization of the weighted likelihood of standard logit models at each iteration. With the increase in the number of classes C, the EM algorithm will get worse due to a need of solving C optimization problems at each iteration. So we take the lower bound approximation of $Q(\boldsymbol{\alpha}_c|\boldsymbol{\psi}^m)$ (see equation 2.12) and the resultant $\boldsymbol{\alpha}_c$ updates are:

$$\boldsymbol{\alpha}_c^{m+1} = \boldsymbol{\alpha}_c^m - [\boldsymbol{B}_{\alpha_c}^m]^{-1}\boldsymbol{g}_{\alpha_c}^m, \tag{2.35}$$

where

$$\boldsymbol{g}_{\alpha_c}^m = \frac{\partial Q(\boldsymbol{\alpha}_c|\boldsymbol{\psi}^m)}{\partial \boldsymbol{\alpha}_c}\bigg|_{\alpha_c=\alpha_c^m} = \sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big[\sum_{t=1}^{T}\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big(d_{itj} - P_{itj}(\boldsymbol{\alpha}_c^m, \boldsymbol{\beta}_{icr})\Big)\Big] \tag{2.36}$$

The Hessian is:

$$\boldsymbol{H}_{\alpha_c}^m = -\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big\{\sum_{t=1}^{T}\Big[\sum_{j=1}^{J}\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}{}^T P_{itj}(\boldsymbol{\alpha}_c^m, \boldsymbol{\beta}_{icr}) - \Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}P_{itj}(\boldsymbol{\alpha}_c^m, \boldsymbol{\beta}_{icr})\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}P_{itj}(\boldsymbol{\alpha}_c^m, \boldsymbol{\beta}_{icr})\Big)^T\Big]\Big\} \tag{2.37}$$

$$\boldsymbol{B}_{\alpha_c}^m = -\frac{1}{2}\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big\{\sum_{t=1}^{T}\Big[\sum_{j=1}^{J}(\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}^T) - \frac{1}{J}\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{itj}\Big)^T\Big]\Big\} \tag{2.38}$$

$$\sum_{r=1}^{R} h_{icr}(\boldsymbol{\beta}_{icr}|\boldsymbol{\psi}^m) = \frac{\sum_{r=1}^{R} s_c^m \mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{icr})}{\sum_{v=1}^{C}\left\{ s_v^m \left[\sum_{r=1}^{R} \mathcal{L}_i(\boldsymbol{\alpha}^m, \boldsymbol{\beta}_{icr}) \right] \right\}} \neq 1$$

$$\implies \boldsymbol{B}_{\alpha_c}^m = \sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr} \boldsymbol{B}_{\alpha_c}^F \tag{2.39}$$

$$\text{where} \quad \boldsymbol{B}_{\alpha_c}^F = -\frac{1}{2}\left\{ \sum_{t=1}^{T}\left[\sum_{j=1}^{J} (\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}^T) - \frac{1}{J}\left(\sum_{j=1}^{J} \boldsymbol{x}_{itj} \right)\left(\sum_{j=1}^{J} \boldsymbol{x}_{itj} \right)^T \right] \right\}.$$

The steps of the MM algorithm to estimate MON-MNL are as follows:

Algorithm 3: MM for the MON-MNL Model

Initialization
Compute $\boldsymbol{B}_{\alpha_c}^F$ using Eq. 2.39 for all $c = 1, \ldots, C$
Initialize parameters $m = 0 : \boldsymbol{\psi}^0 = \{\boldsymbol{\alpha}_1^0, s_1^0, \boldsymbol{\gamma}_1^0, \boldsymbol{\Delta}_1^0, \ldots, \boldsymbol{\alpha}_C^0, s_C^0, \boldsymbol{\gamma}_C^0, \boldsymbol{\Delta}_C^0\}$
while $\|\boldsymbol{\psi}^{m+1} - \boldsymbol{\psi}^m\|_\infty < Tol.$ **do**
  Step 1: Weight $[h_{icr}(.|\boldsymbol{\psi}^m)]$ Calculation ;
  For each i, for each class c, take R draws, with label $\boldsymbol{\beta}_{icr}$, from $\mathcal{N}(\boldsymbol{\gamma}_c^m, \boldsymbol{\Delta}_c^m)$;
  Calculate $\mathcal{L}_i(\boldsymbol{\alpha}_c^m, \boldsymbol{\beta}_{icr})$ for each i using Eq. 2.29;
  Calculate $h_{icr}(.|\boldsymbol{\psi}^m)$ using Eq. 2.31 ;
  Step 2: Update parameters ;
  Update share of each class (s_c^{m+1}) using Eq. 2.32;
  Update $< \boldsymbol{\gamma}_c, \boldsymbol{\Delta}_c >^{m+1}$ for $c = 1, \ldots, C$ using Eq. 2.33;
  Update $\boldsymbol{\alpha}_c^{m+1}$ for $c = 1, \ldots, C$ using Eq. 2.35 ;
end

2.3.4 MON-MNL Estimation using the faster-MM Algorithm

If the number of alternatives is large, similar to LML, the lower bound quadratic approximation in equation 2.38 becomes $\boldsymbol{B}_{\alpha_c}^m = -\frac{1}{2}\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\left\{ \sum_{t=1}^{T}\sum_{j=1}^{J} (\boldsymbol{x}_{itj}\boldsymbol{x}_{itj}^T) \right\}$,

which is a very crude approximation. Therefore, we now extend the faster-MM algorithm to estimate a MON-MNL model with an algorithmic improvement to the update of $\boldsymbol{\alpha}_c$.

Step 1: Compute step length for MM: $\boldsymbol{\mu}^m_{\boldsymbol{\alpha}_c} = -[\boldsymbol{B}^m_{\boldsymbol{\alpha}_c}]^{-1}\boldsymbol{g}^m_{\boldsymbol{\alpha}_c}$.

Step 2: Modify the step length: $\boldsymbol{\zeta}^m_{\boldsymbol{\alpha}_c} = \eta^m_{\boldsymbol{\alpha}_c}\boldsymbol{\mu}^m_{\boldsymbol{\alpha}_c}$.

Step 3: Update $\boldsymbol{\alpha}_c$: $\boldsymbol{\alpha}^{m+1}_c = \boldsymbol{\alpha}^m_c + \boldsymbol{\zeta}^m_{\boldsymbol{\alpha}_c}$.

We compute η^m_γ using:

$$\eta^m_{\boldsymbol{\alpha}_c} = -\frac{(\boldsymbol{\mu}^m_{\boldsymbol{\alpha}_c})^T\boldsymbol{g}^m_{\boldsymbol{\alpha}_c}}{LB}, \tag{2.40}$$

where LB is the lower bound of the quadratic form of the Hessian.[12]

$$LB = -\sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big(\sum_{t=1}^{T} LB^{it}_{\boldsymbol{\alpha}_c}\Big), \tag{2.41}$$

where $LB^{it}_{\boldsymbol{\alpha}_c}$ is a lower bound on the quadratic form of $\boldsymbol{H}^m_{\boldsymbol{\alpha}_c,ijt}$. Equation 2.42 is obtained after rewriting equation 2.37:

$$\boldsymbol{H}^m_{\boldsymbol{\alpha}_c} = \sum_{i=1}^{N}\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big(\sum_{t=1}^{T} \boldsymbol{H}^m_{\boldsymbol{\alpha}_c,ijt}\Big) \tag{2.42}$$

The lower bound $LB^{it}_{\boldsymbol{\alpha}_c}$ is then calculated as follows:

$$\begin{aligned} &LB^{it}_{\boldsymbol{\alpha}_c} = .5(m^2_{it} + M^2_{it} - .5((m_{it} + M_{it})^2)) \\ &\text{where} \quad m_{it} = \min_{1\le j\le J}((\boldsymbol{\mu}^m_{\boldsymbol{\alpha}_c})^T\boldsymbol{x}_{itj} \quad \text{and} \quad M_{it} = \max_{1\le j\le J}((\boldsymbol{\mu}^m_{\boldsymbol{\alpha}_c})^T\boldsymbol{x}_{itj} \end{aligned} \tag{2.43}$$

[12]Note that $\sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m) \neq 1$ for the MON-MNL model.

2.3.5 Standard Errors

Train (2009) and James (2017) suggest to compute the information matrix using cross-product of the M-step scores, which can further be used to obtain standard errors of EM and MM estimates. We derive equations 2.44 and 2.45 to compute simulated scores for $\boldsymbol{\alpha}_c$ and $< \boldsymbol{\gamma}_c, \boldsymbol{\Delta}_c >$, respectively.[13]

$$\frac{\partial Q_i(\boldsymbol{\alpha}_c|\boldsymbol{\psi}^m)}{\partial \boldsymbol{\alpha}_c} = \sum_{r=1}^{R} h_{icr}(.|\boldsymbol{\psi}^m)\Big[\sum_{t=1}^{T}\sum_{j=1}^{J} \boldsymbol{x}_{itj}\Big(d_{itj} - P_{itj}(\boldsymbol{\alpha}_c, \boldsymbol{\beta}_{icr})\Big)\Big] \tag{2.44}$$

$$\begin{aligned}
\frac{\partial\Big[Q_i(\boldsymbol{\gamma}_c, \boldsymbol{\Delta}_c|\boldsymbol{\psi}^m)\Big]}{\partial \boldsymbol{\gamma}_c} &= -\sum_{r=1}^{R}\Big[h_{icr}(.|\boldsymbol{\psi}^m)(\boldsymbol{\Delta}_c)^{-1}(\boldsymbol{\beta}_{icr} - \boldsymbol{\gamma}_c)\Big] \\
\frac{\partial\Big[Q_i(\boldsymbol{\gamma}_c, \boldsymbol{\Delta}_c|\boldsymbol{\psi}^m)\Big]}{\partial \boldsymbol{\Delta}_c} &= \sum_{r=1}^{R}\Big[h_{icr}(.|\boldsymbol{\psi}^m)\Big\{-\frac{1}{2}\boldsymbol{\Delta}_c^{-1} + \frac{1}{2}\boldsymbol{\Delta}_c^{-1}\Big[(\boldsymbol{\beta}_{icr} - \boldsymbol{\gamma}_c)(\boldsymbol{\beta}_{icr} - \boldsymbol{\gamma}_c)^T\Big]\boldsymbol{\Delta}_c^{-1}\Big\}\Big]
\end{aligned} \tag{2.45}$$

In a simulation study, we compared three methods to compute standard errors, namely M-step scores in MM, bootstrapping in MM, and information matrix in MSLE. The standard error estimates of all three methods matched quiet precisely for all parameters, except for some (especially off-diagonal) elements of the variance-covariance matrix (see Table 2.11 in section 2.5.2). However, all standard error estimates of MM bootstrapping and MSLE matched fairly, raising a question on using M-step scores to compute standard errors in EM and MM algorithms.

[13] As an alternative method, bootstrapping also can be used to derive the standard errors, but it is computationally intensive.

We further reviewed existing methods to compute standard errors in EM. Meng and Rubin (1991) suggests to use the SEM algorithm to compute EM standard errors, but SEM is unattractive due to two reasons (Jamshidian and Jennrich, 2000). First, it requires estimation of the Jacobian and Hessian matrices of the M-step objective function, which generally involves cumbersome algebraic operations. Second, SEM is highly sensitive to slower convergence of the algorithm and can result into very high standard errors. In fact, in simulation studies, Jamshidian and Jennrich (2000) and Camilleri (2009) found that methods involving the information matrix of the complete data loglikelihood (e.g., M-step scores and SEM) perform poor in practice, as also verified in our simulation study. They suggest to use the information matrix of the observed (incomplete) loglikelihood at convergence to obtain the correct standard errors of EM. Intuitively, this approach is analogous to switching from EM or MM to Newton-type methods near convergence, but with the motivation to get correct standard errors instead of faster convergence. We suggest to pass the EM or MM estimates to the MSLE with numerical gradient routine. This method does not add any algebraic operations in the original EM or MM algorithm because the loglikelihood is evaluated in the E-step nonetheless. Mathematical simplicity of MM and EM remains intact at the cost of higher total computation time (see section 2.5.2).

2.4 Discussion: advantages and disadvantages of MM over EM and MSLE

James (2017) shows that, unlike EM and MSLE, the Hessian and its inverse need to be computed only once in MM and can be reused at each iteration. This feature makes MM more attractive over other methods when dimensionality of the Hessian is large and inversion is costly. However, we show that this observation is not universally true. For example, computing the inverse of the Hessians $\boldsymbol{B}_{\phi}^{-1}$ and $\boldsymbol{B}_{\alpha}^{-1}$, and reusing them works favorably in MM for LML. However, the Hessian $\boldsymbol{B}_{\alpha_c}^{m}$ and its inverse need to be computed at each iteration, using equation 2.39, in MM for MON-MNL. Nonetheless, the entire computational advantage of MM is not lost because a computationally-intensive part of the Hessian ($\boldsymbol{B}_{\alpha_c}^{F}$) can still be pre-computed in the initialization step.

Since parameter updates in MM just require sufficient statistics and the sample gradient can be written as the sum (in no specific order) of individual gradients, MM estimation is suitable for parallel computation. Even the E-step of the EM algorithm is suitable for parallelization, but the optimization problem in the M-step requires to store weights and simulation draws of the E-step. The communication overhead and storage of these multi-dimensional arrays in EM neutralize the potential benefits of parallelization. We would also like to note that the Hessian and gradient in MSLE can be broken into an unordered sum over individuals and thus estimation can be parallelized. We illustrate the extent of computation time

savings due to parallelization of MM and MSLE in the Monte Carlo study (section 2.5).

Whereas per iteration time of MM is lower than that of EM, MM generally takes more iterations than EM due to a smaller step size. Computational efficiency of MM then hinges upon the trade-off between extra iterations and per-iteration time savings. In fact, we illustrate in the Monte Carlo study that tightness of the Hessian approximation is a key factor in determining the number of extra iterations in MM. Nevertheless, the faster-MM that we have implemented (section 2.2.4) can alleviate the concern of the poor approximation by augmenting the step size while keeping the simplicity of MM.

Previous comparison studies of iterative optimization estimation (Cherchi and Guevara, 2012; James, 2017) often overlooked the issue of MSLE and MM (or EM) needing to maximize two different objective functions; even common tolerance criteria cannot provide a fair comparison between computation efficiency of both methods. Furthermore, computation time of standard errors in EM and MM is often ignored while comparing with MSLE. This is also not appropriate because standard errors can be directly computed using the estimated information matrix in MSLE, but we argue that standard errors obtained from M-step scores in MM and EM may not be correct and additional computation is required (see section 2.3.5 for details). In the Monte Carlo study, we try different convergence tolerances for MM and also report computation time of standard errors to make a fair comparison between MM and MSLE.

2.5 Monte Carlo Study

We conducted two separate Monte Carlo studies for LML and MON-MNL to compare the proposed estimation algorithms (EM, MM, and faster-MM) with the MSLE. MSLE is commonly implemented with numerical gradients –MSLE-Numerical– especially in the context of statistical packages that allow for very flexible utility specifications, but we also considered MSLE with an analytical expression of the gradient –MSLE-Analytical. In both simulation studies, 10 datasets were first simulated to compare all five algorithms. While EM, MM, and faster-MM maximize the same objective function and resulted in similar loglikelihood at convergence, we picked the fastest among them (faster-MM) and compared it with MSLE-Analytical in a detailed Monte Carlo study. Using 40 simulated datasets, we compared these methods based on various performance metrics, namely: log-likelihood at the convergence, parallel and sequential estimation time, average percentage bias (APB), finite sample standard error (FSSE), and percentage difference between finite and asymptotic standard errors[14] (StdFssePerDiff). The sensitivity of performance metrics was evaluated relative to the number of simulation draws $\{300, 500, 1000\}$, the convergence tolerance of faster-MM[15] $\{10^{-3}, 10^{-4}, 10^{-6}\}$, the sample size $\{500, 2000\}$ (with 5 choice situations), and the number of alterna-

[14]Since standard errors estimation in LML requires bootstrapping, FSSE and StdFssePerDiff were not estimated due to computational constraints.

[15]We kept the standard tolerance criterion for MSLE: lower bound of 10^{-6} on the change in the value of the objective function during a step.

tives $\{4, 50, 100\}$[16]. In terms of variation in the specification of unobserved preference heterogeneity, 2^{nd} and 4^{th} order polynomials – $\boldsymbol{y}(\boldsymbol{\beta}_{ir})$ in equation (2.4) – were considered in LML. Since a one-class MON-MNL is the same as a parametric mixed MNL (MMNL) model, one- and two-class specifications were considered. We coded all the estimation methods in MATLAB. Code is available upon request.

Definitions of performance metrics for a scalar parameter are given below. For succinctness, we reported averages across all parameters.

APB: The parameters across all simulated datasets are averaged and the subsequent, mean parameter estimate is used to compute APB:

$$APB = \left| \frac{\text{Mean Parameter Estimate - True Parameter Value}}{\text{True Parameter Value}} \right| \times 100.$$

FSSE: The standard deviation of the parameter estimates across the simulated data sets.

StdFssePerDiff: The asymptotic standard errors across all simulated datasets are averaged, and then the mean asymptotic standard error and FSSE are used to compute this metric:

$$\text{StdFssePerDiff} = \left| \frac{\text{Mean Asymptotic Standard Error - FSSE}}{\text{Mean Asymptotic Standard Error}} \right| \times 100.$$

[16]For LML, we kept 4 alternatives in all simulations. A variation in cardinality of the choice set was only considered for MON-MNL to evaluate the extant of an eventual poor approximation in MM and also an improvement due to faster-MM.

2.5.1 LML Monte Carlo Study

Simulation Setup

We considered a data generating process (DGP), where the indirect utility of a decision-maker i for alternative j in choice situation t is: $U_{itj} = x_{1itj}\alpha_1 + x_{2itj}\alpha_2 + z_{1itj}\beta_{1i} + z_{2itj}\beta_{2i} + \varepsilon_{itj}$. The attributes x_{1itj}, x_{2itj}, z_{1itj}, and z_{2itj} were drawn from a standard normal distribution and ε_{itj} was drawn from a standard Type-I Extreme Value distribution. α_1 and α_2 are fixed coefficients. β_{1i} and β_{2i} are random parameters with bimodal-normal distribution.

In LML, the number of dimensions of the support set S (Equation 2.4) is equal to the number of random parameters ($RP = 2$). Boundaries of the support set S were defined considering 3 standard deviation away from the true mean of the random parameters. Subsequently, each dimension was divided into 10^3 equally-spaced points, leading to a multidimensional grid that contains $10^{(3\times RP)}$ points. We take random draws for each person from the grid to compute the simulate loglikelihood.[17] LML estimates histograms of random parameters, but we report mean and standard deviation of the histograms as parameters of interest.

[17] See Train (2016) for a detailed discussion about setting boundaries of the support set and sampling from the grid.

Results and Discussion

The average (across the 10 repetitions) loglikelihood, estimation time, and number of iterations are presented in Table 2.1. The loglikelihood values of all methods are virtually indistinguishable.

Poor computational performance of MM is evident. This might have happened because in the MM algorithm, the curvature loses its sharpness due to a poor Hessian approximation, meaning that MM might have stuck in a specific region of the loglikelihood due to a small step size and poor search direction.

The proposed faster-MM algorithm addressed the poor Hessian approximation of MM and considerably improved computational efficiency. The faster-MM algorithm reduced MM estimation time by a factor of $\frac{1}{103}$, and even outperformed EM and MSLE-Numerical. Note that time-per-iteration of MM and faster-MM are very close, which are 0.270 seconds and 0.278 seconds. This small difference is expected because the faster-MM algorithm only needs some additional elementary math operations relative to MM to compute the scalar vector η_{ϕ}^{m} to augment the step size (section 2.2.4).

As clearly shown in Table 2.1, MSLE-Analytical is computationally superior for LML estimation. Even though MSLE-Analytical and MSLE-Numerical need virtually the same number of iterations, the latter requires more loglikelihood evaluations. Nonetheless, in this simulation MSLE-Numerical appears to be the second best method, after MSLE-Analytical, to estimate the LML model.

Table 2.1: Preliminary Monte Carlo Simulation Results (LML Model)

	Loglikelihood	**Estimation Time (min)**	**Iterations**
EM	-1987.4678	1.06	79
MM	-1987.4289	66.18	14698
Faster MM	-1987.4255	0.64	138
MSLE-Numerical	-1987.7870	0.53	44
MSLE-Analytical	-1987.7866	0.13	45
Tolerance		1.00E-03	
# of Observations		500	
# of Draws		500	
Polynomial Order		2	

Tables 2.2 to 2.4 summarize results of a detailed Monte Carlo study which compares faster-MM and MSLE-Analytical based on various performance metrics. Scenarios 1, 2, and 3 correspond to simulation draws of 300, 500, and 1000, and additional indices (a), (b), and (c) characterize the faster-MM tolerance criteria of 10^{-3}, 10^{-4}, and 10^{-6}. This nomenclature remains the same for the MON-MNL Monte Carlo study (Tables 2.7-2.10 in section 2.5.2).

The stricter tolerance criterion in faster-MM may not result into any significant improvement in model fit (i.e., loglikelihood) and APB, but increases estimation time significantly. For instance, whereas scenarios 1(a) and 1(c) in Table 2.3 (N=500, polynomial order = 4) attain similar loglikelihood and APB, parallel estimation time ('PET') of scenario 1(c) is approximately 11.5 times that of 1(a).

In contrast to intuition, irrespective of the estimation method, increase in simulation draws in LML does not necessarily improve model fit but generally in-

creases estimation time with no visible effect on APB. For instance, comparing scenarios 2 and 3 in Table 2.2 illustrates that MSLE-Analytical with 1000 draws converged to a lower loglikelihood than with 500 draws, with twice the estimation time.

Even though parallelization of faster-MM on 8 MATLAB workers could reduce computation time by around 65% to 75%, MSLE-Analytical remains superior because parallel estimation also reduced computation time of MSLE by around 55% to 65% (see columns 'SET', 'PET', and 'SET/PET' in Tables 2.2-2.4).

The step size multiplier ('shrinkage') η_{ϕ}^{m} in faster-MM appears to increase monotonically with the number of draws across all cases. This is because the lower-bound Hessian approximation of MM gets poorer (i.e., flatter curvature) with the increase in simulation draws and thus, faster-MM has more scope to improve the step-size and convergence (see section 2.2.4). Lower values of shrinkage in Table 2.3 (polynomial order 4) relative to Table 2.2 (polynomial order 2) indicate possibilities of poorer performance of faster-MM with the increase in flexibility. However, sample size may not have any effect on the ability of faster-MM to improve the MM approximation (compare 'shrinkage' column in Tables 2.2 and 2.4).

Table 2.2: Monte Carlo Simulation Results (LML Model, N=500, Polynomial Order=2)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET [a]	SET [b]	SET/PET	APB [c]	Shrinkage
300	-	MSLE (A) [d]	1	-2015.791	0.19	0.43	2.3	27.3	-
500	-	MSLE (A)	2	-1987.787	0.27	0.75	2.8	30.1	-
1000	-	MSLE (A)	3	-2001.552	0.52	1.38	2.6	27.8	-
300	1.00E-03	Faster MM	1(a)	-2015.629	0.66	2.11	3.2	28.1	30.4
500	1.00E-03	Faster MM	2(a)	-1987.645	1.01	3.58	3.6	30.8	50.2
1000	1.00E-03	Faster MM	3(a)	-2001.462	2.12	7.10	3.3	28.7	94.1
300	1.00E-04	Faster MM	1(b)	-2015.609	0.95	2.99	3.1	27.5	29.9
500	1.00E-04	Faster MM	2(b)	-1987.623	1.46	5.21	3.6	30.2	50.5
1000	1.00E-04	Faster MM	3(b)	-2001.439	3.03	9.99	3.3	28.0	97.1
300	1.00E-06	Faster MM	1(c)	-2015.607	1.49	4.57	3.1	27.3	30.6
500	1.00E-06	Faster MM	2(c)	-1987.621	2.37	8.23	3.5	30.0	49.2
1000	1.00E-06	Faster MM	3(c)	-2001.436	4.93	16.28	3.3	27.8	94.3

[a] Parallel estimation time (in minutes)

[b] Sequential estimation time (in minutes)

[c] Mean of average percentage bias of all parameters

[d] MSLE (analytical gradient)

Table 2.3: Monte Carlo Simulation Results (LML Model, N=500, Polynomial Order=4)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET	SET	SET/PET	APB	Shrinkage
300	-	MSLE (A)	1	-2003.748	0.70	1.71	2.4	27.0	-
500	-	MSLE (A)	2	-1975.715	1.13	2.71	2.4	28.5	-
1000	-	MSLE (A)	3	-1988.064	2.30	5.53	2.4	25.4	-
300	1.00E-03	FasterMM	1(a)	-2004.419	2.58	8.04	3.1	29.8	21.7
500	1.00E-03	FasterMM	2(a)	-1976.657	10.83	32.65	3.0	28.3	38.9
1000	1.00E-03	FasterMM	3(a)	-1988.077	28.72	87.91	3.1	28.9	58.8
300	1.00E-04	FasterMM	1(b)	-2003.732	7.89	24.50	3.1	28.1	21.5
500	1.00E-04	FasterMM	2(b)	-1975.900	26.56	83.57	3.1	29.8	39.7
300	1.00E-06	FasterMM	1(c)	-2003.586	29.58	90.97	3.1	27.5	21.5

Note: Refer Table 2.2 for description of column headers

Table 2.4: Monte Carlo Simulation Results (LML Model, N=2000, Polynomial Order=2)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET	SET	SET/PET	APB	Shrinkage
300	-	MSLE (A)	1	-8069.076	1.77	4.75	2.7	29.2	-
500	-	MSLE (A)	2	-8028.186	3.17	8.54	2.7	30.2	-
1000	-	MSLE (A)	3	-8018.608	6.79	18.67	2.8	29.2	-
300	1.00E-03	FasterMM	1(a)	-8069.209	7.15	25.99	3.6	29.4	30.1
500	1.00E-03	FasterMM	2(a)	-8028.369	14.54	52.55	3.6	30.5	49.1
1000	1.00E-03	FasterMM	3(a)	-8018.703	34.49	130.59	3.8	29.8	94.3
300	1.00E-04	FasterMM	1(b)	-8068.990	9.55	34.40	3.6	29.2	30.0
500	1.00E-04	FasterMM	2(b)	-8027.806	19.42	71.80	3.7	30.5	48.9
300	1.00E-06	FasterMM	1(c)	-8068.487	14.52	52.31	3.6	29.0	30.7

Note: Refer Table 2.2 for description of column headers

2.5.2 MON-MNL Monte Carlo Study

Simulation Setup

We considered a data generating process (DGP) with 2 classes (with proportion $s_2 = .7$), where utility of individual i of class c from choosing alternative j at time t is:

$$U^c_{itj} = x_{1itj}\alpha_{c1} + x_{2itj}\alpha_{c2} + z_{1itj}\beta^{c1}_i + z_{2itj}\beta^{c2}_i + \varepsilon_{itj} \tag{2.46}$$

where α_{c1} and α_{c2} are fixed parameters, and β^{c1}_i and β^{c2}_i are random parameters of class c with the following distribution:

$$\begin{pmatrix}\beta^{c1}_i \\ \beta^{c2}_i\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\gamma_{c1} \\ \gamma_{c2}\end{pmatrix}\begin{pmatrix}\Delta^{11}_c & \Delta^{12}_c \\ \Delta^{21}_c & \Delta^{22}_c\end{pmatrix}\right). \tag{2.47}$$

All other specifications remain the same as of LML DGP (see section 2.5.1). For one-class MON-MNL (i.e., a mixed logit MMNL), datasets were also generated using equations 2.46 and 2.47, but a single class DGP was considered.

Results and Discussion

Tables 2.5 and 2.6 summarize the average (across 10 DGPs) loglikelihood, estimation time (in minutes), and the number of iterations for MMNL (one class) and

MON-MNL (two classes). Across both choice set cardinalities and model specifications, all three EM-based algorithms converged to the same loglikelihood value (up to 3 decimal places), which is marginally different than that of MSLE. This minor discrepancy might be a result of the different treatment of simulation error while maximizing a different objective function (in EM and quasi-Newton).

Resemblance between per iteration time of MM and faster-MM, as also pointed out in the LML Monte Carlo study, holds for MMNL and MON-MNL models. For a small choice set with $J = 4$, under the MMNL specification, MM and faster-MM took similar estimation time and also outperformed all other methods by a significant margin. While the latter observation is consistent with James (2017), the former resemblance can be attributed to a tighter approximation of Hessian in MM for a small choice set (i.e., shrinkage value $\eta^m_{\alpha_c}$ of just 1.15, see note in Table 2.5).

Under the MON-MNL specification, with a small choice set $J = 4$, MM and faster-MM could still computationally surpass EM and MSLE-Numerical, but MSLE-Analytical outperformed all methods. In fact, under both MMNL and MON-MNL specifications, the performance of MM estimation gets worse for a choice set with higher cardinality $J = 50$. We argue that the computational superiority of MM, as noted by James (2017), seems to be true for a narrow set of logit specifications and settings.

Due to a Hessian approximation that deteriorates, the value of the scalar multiplier $\eta^m_{\alpha_c}$ in faster-MM increases with the number of alternatives. The shrinkage

$\eta^m_{\alpha_c}$ effectively attained average values of 4.78 and 4.73 for $J = 50$ in MMNL and MON-MNL (see note in Tables 2.5 and 2.6), which led to significant computation time savings of the faster-MM against MM – the ratio of estimation time of MM against faster-MM is 1.95 and 4.24, respectively. Whereas the faster-MM algorithm was indeed able to significantly reduce MM estimation time for larger choice sets, MSLE-Analytical marginally outperformed it under both specifications.

Table 2.5: Preliminary Monte Carlo Simulation Results (MMNL Model)

Alternatives	EM	MM	Faster MM	MSLE-Numerical	MSLE-Analytical
Loglikelihood					
4	-8890.3675	-8890.3678	-8890.3678	-8889.7890	-8889.7891
50	-18902.6051	-18902.6059	-18902.6055	-18897.9788	-18897.9793
Iterations					
4	28	33	36	20	21
50	87	215	104	19	22
Estimation time (min)					
4	1.35	0.38	0.33	1.65	0.99
50	246.50	130.96	66.91	132.48	32.59
Tolerance	1.00E-04				
# of Observations	2000				
# of Draws	500				

Note: Average shrinkage values of faster-MM for $J = 4$ and $J = 50$ are 1.15 and 4.78, respectively.

Tables 2.7 to 2.10 summarize the average (across 40 simulated datasets) of performance metrics for a detailed comparison of MON-MNL estimation using MSLE-Analytical and faster-MM.[18] As expected, increase in cardinality of the choice set increased the bias in estimates (compare 'APB' in Tables 2.7 and 2.8), but there is no noticeable difference between APB of faster-MM and MSLE-Analytical.

[18]The results of the detailed Monte Carlo study for MMNL did not provide additional insights and therefore, those results are not presented for brevity.

Table 2.6: Preliminary Monte Carlo Simulation Results (MON-MNL Model)

Alternatives	EM	MM	Faster MM	MSLE-Numerical	MSLE-Analytical
Loglikelihood					
4	-7958.2938	-7958.2933	-7958.2933	-7957.0813	-7957.0888
50	-15174.4599	-15174.4593	-15174.4594	-15172.9390	-15172.9348
Iterations					
4	92	167	141	44	45
50	61	921	208	61	59
Estimation time (min)					
4	16.3	8.8	7.5	9.7	4.6
50	171.3	362.8	85.4	119.6	53.6
Tolerance	1.00E-04				
# of Observations	2000				
# of Draws	500				
# of Classes	2				

Note: Average shrinkage values of faster-MM for $J = 4$ and $J = 50$ are 1.12 and 4.73, respectively.

In fact, FSSE and StdFssePerDiff of both estimation methods remain virtually same across all four combinations of sample size and choice set cardinality. The sensitivity of performance metrics relative to simulation draws and tolerance criterion (in faster-MM) remains consistent with the results of the LML simulation (see section 2.5.1). Whereas parallel computation could reduce estimation time of both methods by around 67% to 78% (see 'SET/PET' column in Tables 2.7-2.10), MSLE-Analytical still outperformed faster-MM. However, whereas the ratio of estimation time of both methods remained intact, the scale of difference in estimation time decreased significantly under parallelization. For instance, for a sample size of 500 and choice set with 100 alternatives, comparing scenarios 2 and 2(b) in Table 2.10) shows that the difference in estimation time of MSLE-Analytical and faster-MM reduced from 57.94 minutes in sequential ('SET') to 12.09 minutes in

parallel ('PET') computation.

As detailed in section 2.3.5, whereas MSLE-Analytical could retrieve standard errors as a by-product of estimation, we had to compute standard errors of faster-MM by passing point estimates through MSLE-Numerical.[19] Tables 2.7 to 2.10 show that the standard error computation ('SST') may often take as much time as point estimation ('SET'). However, since MSLE-Numerical could take advantage of the benefits of parallel computation, parallelization could reduce this overhead time by around 60% to 80%.

[19]We could have computed analytical scores, and thus Hessian and standard errors, passing faster-MM estimates to the score function of MSLE-Analytical. This may not add extra computation time to faster-MM, but computation time comparison would not be fair (see section 2.3.5).

Table 2.7: Monte Carlo Simulation Results (MON-MNL Model, N=2000, J=4)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET [a]	PST [b]	SET [c]	SST [d]	SET/PET	APB [e]	FSSE [f]	StdFssePerDiff [g]
300	-	MSLE (A)	1	-7952.133	0.66	-	2.70	-	4.1	14.4	0.0515	17.0
500	-	MSLE (A)	2	-7949.889	1.08	-	4.11	-	3.8	14.0	0.0486	19.5
1000	-	MSLE (A)	3	-7948.890	2.16	-	7.23	-	3.3	14.8	0.0499	17.7
300	1.00E-03	Faster MM	1(a)	-7952.737	1.06	1.24	3.44	3.46	3.3	17.3	0.0525	18.9
500	1.00E-03	Faster MM	2(a)	-7950.299	1.95	2.01	6.75	5.59	3.5	17.8	0.0474	21.2
300	1.00E-04	Faster MM	1(b)	-7952.730	1.10	1.23	3.45	3.48	3.1	17.2	0.0528	18.7
500	1.00E-04	Faster MM	2(b)	-7950.293	2.05	1.98	6.97	5.62	3.4	15.2	0.0475	21.1
1000	1.00E-04	Faster MM	3(b)	-7949.108	5.58	3.72	18.68	10.70	3.4	14.5	0.0495	18.3
500	1.00E-06	Faster MM	2(c)	-7950.293	2.11	2.06	7.46	5.66	3.5	15.2	0.0475	21.0

[a] Parallel estimation time (in minutes)

[b] Parallel standard error computation time (in minutes)

[c] Sequential estimation time (in minutes)

[d] Sequential standard error computation time (in minutes)

[e] Mean of average percentage bias of all parameters

[f] Mean of finite sample standard error of all parameters

[g] Mean of percentage difference in finite sample and asymptotic standard error of all parameters

[h] MSLE (analytical gradient)

Table 2.8: Monte Carlo Simulation Results (MON-MNL Model, N=2000, J=50)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET	PST	SET	SST	SET/PET	APB	FSSE	StdFssePerDiff
300	-	MSLE (A)	1	-15184.466	9.80	-	35.55	-	3.6	63.2	0.0457	11.4
500	-	MSLE (A)	2	-15176.935	16.24	-	56.40	-	3.5	65.7	0.0455	12.2
1000	-	MSLE (A)	3	-15167.718	34.40	-	134.01	-	3.9	63.1	0.0451	9.1
300	1.00E-03	Faster MM	1(a)	-15188.429	11.31	7.18	41.43	34.84	3.7	61.4	0.0509	15.7
500	1.00E-03	Faster MM	2(a)	-15178.465	19.15	12.95	81.00	58.11	4.2	61.8	0.0475	12.2
300	1.00E-04	Faster MM	1(b)	-15188.426	11.66	7.52	42.26	34.16	3.6	61.3	0.0508	15.5
500	1.00E-04	Faster MM	2(b)	-15178.459	19.92	12.93	84.29	58.16	4.2	61.7	0.0473	12.0
1000	1.00E-04	Faster MM	3(b)	-15168.587	67.63	33.12	293.57	157.80	4.3	65.3	0.0477	13.1
500	1.00E-06	Faster MM	2(c)	-15178.459	20.17	12.61	84.85	57.98	4.2	61.6	0.0472	12.0

Note: Refer Table 2.7 for description of column headers

Table 2.9: Monte Carlo Simulation Results (MON-MNL Model, N=500, J=50)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET	PST	SET	SST	SET/PET	APB	FSSE	StdFssePerDiff
300	-	MSLE (A)	1	-3757.109	3.52	-	12.57	-	3.6	60.7	0.0987	24.1
500	-	MSLE (A)	2	-3756.219	5.05	-	19.16	-	3.8	61.6	0.0990	22.4
1000	-	MSLE (A)	3	-3753.961	11.62	-	44.86	-	3.9	60.8	0.0951	19.2
300	1.00E-03	Faster MM	1(a)	-3758.696	4.15	3.30	18.59	15.41	4.5	61.7	0.0981	20.8
500	1.00E-03	Faster MM	2(a)	-3757.204	8.53	5.38	39.04	22.92	4.6	61.8	0.1020	24.8
300	1.00E-04	Faster MM	1(b)	-3758.688	4.32	3.67	19.06	16.41	4.4	64.3	0.0978	20.9
500	1.00E-04	Faster MM	2(b)	-3757.195	9.20	5.49	42.24	24.99	4.6	61.5	0.1018	24.9
1000	1.00E-04	Faster MM	3(b)	-3754.389	21.64	10.83	100.76	49.28	4.7	61.1	0.0951	20.7
500	1.00E-06	Faster MM	2(c)	-3757.194	9.90	5.43	44.49	22.40	4.5	61.4	0.1017	24.9

Note: Refer Table 2.7 for description of column headers

Table 2.10: Monte Carlo Simulation Results (MON-MNL Model, N=500, J=100)

Draws	Tolerance	Algorithm	Scenarios	Loglikelihood	PET	PST	SET	SST	SET/PET	APB	FSSE	StdFssePerDiff
300	-	MSLE (A)	1	-4153.599	3.45	-	15.46	-	4.5	76.5	0.1015	24.8
500	-	MSLE (A)	2	-4148.869	5.62	-	24.33	-	4.3	75.7	0.0986	25.1
1000	-	MSLE (A)	3	-4148.613	12.20	-	52.20	-	4.3	76.1	0.0986	24.9
300	1.00E-03	Faster MM	1(a)	-4156.026	5.39	3.72	22.16	17.63	4.1	74.6	0.0975	23.7
500	1.00E-03	Faster MM	2(a)	-4150.455	14.09	6.32	64.08	29.84	4.5	75.5	0.0999	23.0
300	1.00E-04	Faster MM	1(b)	-4156.020	5.84	4.15	23.77	17.62	4.1	74.3	0.0972	23.8
500	1.00E-04	Faster MM	2(b)	-4150.446	17.71	6.24	82.27	29.72	4.6	75.3	0.0998	22.9
1000	1.00E-04	Faster MM	3(b)	-4149.348	23.55	11.75	106.84	54.50	4.5	75.6	0.0981	22.0
500	1.00E-06	Faster MM	2(c)	-4150.445	25.78	6.52	119.50	29.97	4.6	75.2	0.0998	22.9

Note: Refer Table 2.7 for description of column headers

Table 2.11: Standard Errors Comparison

	Data Generating Process 1				Data Generating Process 2			
	faster-MM			MSLE-Analytical	faster-MM			MSLE-Analytical
Class 1	**M-step Scores**	**Bootstrap**	**MM-MSLE (Corrected)**	**Information Matrix**	**M-step Scores**	**Bootstrap**	**MM-MSLE (Corrected)**	**Information Matrix**
α_{11}	0.06	0.07	0.06	0.06	0.14	0.12	0.13	0.12
α_{12}	0.09	0.09	0.08	0.08	0.18	0.21	0.19	0.18
γ_{11}	0.14	0.13	0.13	0.13	0.24	0.27	0.28	0.23
γ_{12}	0.17	0.20	0.16	0.17	0.31	0.34	0.32	0.31
Δ_1^{11}	0.38	0.37	0.35	0.36	0.41	0.34	0.35	0.34
Δ_1^{12}	**0.60**	0.27	0.24	0.26	**0.71**	0.37	0.39	0.33
Δ_1^{22}	0.60	0.61	0.53	0.56	**0.81**	0.67	0.71	0.69
Class 2								
α_{21}	0.07	0.09	0.07	0.07	0.08	0.08	0.09	0.08
α_{22}	0.09	0.13	0.10	0.10	0.13	0.12	0.12	0.12
γ_{21}	0.12	0.17	0.13	0.13	0.15	0.15	0.16	0.16
γ_{22}	0.16	0.17	0.17	0.17	0.21	0.19	0.19	0.21
Δ_2^{11}	0.27	0.32	0.25	0.26	0.41	0.42	0.44	0.40
Δ_2^{12}	**0.38**	0.19	0.18	0.18	**0.51**	0.29	0.30	0.27
Δ_2^{22}	0.34	0.35	0.32	0.32	0.49	0.46	0.42	0.48
Loglikelihood	-2088.8			-2089.3	-1549.7			-1549.3
N	500							
J	4							

Note: The discrepant standard errors are in **bold**-font.

Table 2.11 illustrates the discrepancies in the faster-MM standard errors, which are computed using M-step scores under the two DGPs (differing in true parameter values). A considerable deviation was mainly exhibited in off-diagonal terms of the variance-covariance matrices Δ_1^{12} and Δ_2^{12}. However, the corrected MM-MSLE method could closely mimic the standard errors derived from both bootstrapping and MSLE-Analytical.

2.6 Empirical Study: Adoption of Electric Motorcycles

The empirical study relies on microdata collected from a stated preference survey of consumers' willingness to adopt electric motorcycles in Solo, Indonesia in August and September, 2015. Lower cost, lower performance e-bikes and electric scooters exist in Solo but have very low market penetration. The purpose of the survey was to determine whether, at what price point, and at what quality, consumers would replace gas-powered motorcycles, which contribute to elevated levels of local pollution, for electric motorcycles. Guerra (2017) provides details on the survey design, survey collection, and data processing.

In this study we rely on the data from the 1208 respondents (out of 1307), who completed all choice scenarios (5 in total). In each choice scenario, respondents selected among a conventional motorcycle, an electric motorcycle, and no motorcycle based on price, speed, range, and charge time. The relevant attributes and their levels are presented in Table 2.12. Purchase price is a monthly credit pay-

ment of 500,000 rupiah (roughly US$40) for three years. In Solo as in much of Indonesia, consumers buy motorcycles on credit and these are the most common repayment terms.

Table 2.12: Summary of choice sets and attribute values

Attributes	**Gas Motorcycle**	**Electric Motorcycle**
Monthly payment (3 years in thousands)	500	400, 500, 600
Fuel price (thousands of rupiah per liter)	7, 8, 9	4, 5, 6, 7, 8
Maximum range on single charge (km)	NA	40, 60, 80, 100
How long to charge (hours)	NA	2, 3, 4, 5
Max speed (km/h)	100	60, 70, 80, 90, 100

Table 2.13 presents the summary statistics for the 1208 respondents. Comparing the sample to Solo's resident population, the sample is younger, more male, and includes slightly fewer of the lowest income households. It is also a group that is more likely to be making future motorcycle purchases.

Since the focus of this study is to compare various estimation methods, we have specified a model in which the utility equation has alternative-specific attributes. This empirical specification can be further refined by exploring heterogeneity across different demographic groups (Jones et al., 2013; Guerra, 2017). We estimated MMNL (one class), MON-MNL (two classes), and LML models using faster-MM and MSLE-Analytical in preference space[20]. LML estimation accuracy was validated by comparing estimates of MMNL with a second-

[20]Although the scale of the parameter estimates varies by model, dividing motorcycle feature estimates by negative of fuel price coefficient produces consistent WTP estimates.

Table 2.13: Descriptive sample statistics (N=1208)

Statistic	Mean	St. Dev.	Min	Max
Woman	0.39	0.49	0	1
Age	32	12.6	16	80
Household motorcycles	2.11	1.07	0	8
Has own motorcycle	0.95	0.21	0	1
Motorcycle: principal mode	0.81	0.39	0	1
Monthly income (100,000's of Rupiah)				
<500	0.07	0.25	0	1
500-1,000	0.31	0.46	0	1
1,000-2,500	0.41	0.49	0	1
2,500-5,000	0.17	0.38	0	1
<5,000	0.04	0.19	0	1

order LML-Polynomial specification –i.e., $\boldsymbol{y}(\boldsymbol{\beta}_{ir})$ to be a second-order polynomial– since both specifications are analytically equivalent. To highlight the flexibility of LML and its importance in modeling multimodality, an eighth-order LML-polynomial model was also estimated. The boundaries of the parameter space were set to three standard deviations away from the estimated mean of MMNL. In MMNL and MON-MNL, correlations across parameters and standard deviation of monthly payment and fuel price are assumed to be zero since they were not statistically significant when included. We also analyzed the sensitivity of performance metrics – computation time, parameter estimates, and model fit – relative to simulation draws and faster-MM tolerance criteria.

Tables 2.14, 2.15, and 2.16 summarize estimation results. From a behavioral perspective, some new insights are unveiled. Across all three models, respondents

are more likely to choose a motorcycle with higher speed, longer range, faster charging times, lower fuel cost, and lower monthly payment. This is both intuitive and consistent with the literature on alternative fuel vehicles (Cherry and Cervero, 2007; Daziano, 2013; Jones et al., 2013). Charge time is particularly important with respondents indicating a WTP of a 20% fuel premium for an hour shorter charge time. There is substantial variation in how much respondents value speed, range, and charge times. For example, across the MMNL and LML models the standard deviation of the parameter estimates for speed, range, and charger time are larger than the point estimates.

The results of this empirical application are aligned with those obtained in the Monte Carlo study, but additional remarks are uncovered. Comparing scenarios 1(a) and 1(b) across all models indicates that a tolerance criterion of $10^{(-3)}$ in faster-MM is strict enough to obtain stable estimates and model fit that are comparable with MSLE-Analytical. Furthermore, scenario 2 (with higher simulation draws) resulted into slightly better fit than scenario 1 –1(a) in faster-MM– across all models, but at the expense of around three- to five-fold computational time increase with no noticeable differences in parameter estimates. The standard error estimates of faster-MM (scenario 1(a)) and MSLE-Analytical (scenario 1) did not exhibit any considerable difference across all three models. Parallel computation could reduce estimation time by 50% to 80% which is consistent with the simulation findings.

As expected, estimated parameters, standard errors, and loglikelihood values

of MMNL and LML-Polynomial of second-order are similar. The eighth-order LML-Polynomial has a higher loglikelihood value relative to the second-order LML-Polynomial due to a higher permitted by more parameters. However, comparing scenario 3 in Table 2.15 and scenario 2 in Table 2.16 shows that MON-MNL could attain a much higher loglikelihood (-2615.88) with fewer parameters than the eighth-order LML-Polynomial (-2687.80).

Table 2.14: Adoption of Electric Motrocycles in Indonesia (MMNL Model Results)

	Est.	**t-stat**	**Est.**	**Est.**	**t-stat**	**Est.**	**Est.**
ASC (Electric)	8.73	9.36	8.50	7.92	9.35	8.13	7.43
ASC (Gas)	11.23	11.22	11.04	10.04	9.19	10.24	9.89
Monthly pay (Rp. millions)	-2.73	-3.06	-2.78	-2.71	-3.20	-2.71	-2.68
Fuel price (Rp. thousands)	-0.68	-12.85	-0.69	-0.68	-12.60	-0.68	-0.66
Mean							
Charge time (hours)	-0.92	-8.21	-1.09	-1.16	-9.59	-1.17	-1.09
Max speed(km/h/100)	4.60	7.08	4.66	4.97	7.34	4.87	4.81
Max range(km/100)	1.94	4.08	2.40	2.28	4.95	2.26	2.59
Standard Deviation							
Charge time (hours)	1.37	10.06	1.50	1.51	12.30	1.51	1.49
Max speed(km/h/100)	6.42	11.92	6.10	5.63	10.90	5.69	5.63
Max range(km/100)	5.64	11.34	5.17	4.80	9.10	4.79	4.58
Loglikelihood	-2721.54		-2717.32	-2722.14		-2722.05	-2719.31
Parallel est. time (min)	1.15		3.41	1.13		1.28	4.19
Sequential est. time (min)	2.15		6.84	2.19		2.55	10.01
Number of Draws	300		1000	300		300	1000
Tolerance	-		-	1.00E-03		1.00E-04	1.00E-03
Scenario	1		2	1(a)		1(b)	2
Method	MSLE-Analytical			Faster MM			

Table 2.15: Adoption of Electric Motrocycles in Indonesia (LML Model Results)

	Est.	**t-stat**	**Est.**	**Est.**	**Est.**	**t-stat**	**Est.**	**Est.**
ASC (Electric)	12.35	8.82	10.93	9.67	12.09	8.41	12.27	10.74
ASC (Gas)	14.59	11.36	13.29	12.37	14.27	10.95	14.49	13.10
Monthly pay (Rp. millions)	-3.31	-2.66	-2.98	-2.92	-3.29	-2.56	-3.31	-2.99
Fuel price (Rp. thousands)	-0.71	-11.72	-0.69	-0.74	-0.71	-11.95	-0.71	-0.70
Mean								
Charge time (hours)	-1.15	-7.68	-1.10	-0.97	-1.17	-7.38	-1.16	-1.12
Max speed(km/h/100)	3.91	6.96	3.61	4.83	4.05	7.16	3.94	3.85
Max range(km/100)	1.75	4.28	1.87	1.76	1.71	4.13	1.74	1.88
Standard Deviation								
Charge time (hours)	1.44	9.34	1.40	1.32	1.47	8.53	1.45	1.43
Max speed(km/h/100)	7.13	9.72	6.24	8.18	6.98	9.24	7.07	6.20
Max range(km/100)	5.36	10.04	5.13	5.48	5.50	10.21	5.40	5.27
Loglikelihood	-2729.02		-2724.02	-2687.80	-2729.25		-2729.04	-2724.22
Parallel est. time (min)	1.7		7.6	34.9	5.9		8.3	26.4
Sequential est. time (min)	3.5		15.2	76.6	16.2		22.5	76.6
Number of Draws	300		1000	1000	300		300	1000
Order	2		2	8	2		2	2
Tolerance	-		-	-	1.00E-03		1.00E-04	1.00E-03
Scenario	1		2	3	1(a)		1(b)	2
Method	MSLE-Analytical				Faster MM			

Table 2.16: Adoption of Electric Motrocycles in Indonesia (MON-MNL Model Results)

	Class 1		Class 2		Class 1	Class 2	Class 1		Class 2		Class 1	Class 2	Class 1	Class 2
	Est.	**t-stat**	**Est.**	**t-stat**	**Est.**	**Est.**	**Est.**	**t-stat**	**Est.**	**t-stat**	**Est.**	**Est.**	**Est.**	**Est.**
ASC (Electric)	14.60	13.88	13.30	10.31	18.48	15.33	14.67	14.22	12.98	9.70	16.97	16.22	16.97	16.22
ASC (Gas)	62.19	4.30	15.06	11.49	67.57	17.18	65.91	4.65	14.91	10.80	68.44	17.36	68.44	17.36
Monthly pay (Rp. millions)	79.91	3.94	-5.76	-5.26	74.55	-5.82	72.30	4.14	-5.61	-5.10	81.58	-5.55	81.58	-5.55
Fuel price (Rp. thousands)	-23.58	-4.30	-0.12	-1.67	-24.89	-0.16	-25.18	-4.19	-0.12	-1.74	-26.78	-0.15	-26.78	-0.15
Mean														
Charge time (hours)	-10.58	-4.10	-0.89	-5.44	-9.94	-0.93	-11.16	-4.43	-0.82	-5.70	-10.34	-1.02	-10.34	-1.02
Max speed(km/h/100)	145.08	4.27	2.48	2.49	138.74	2.69	152.38	4.69	2.59	2.68	134.95	2.93	134.95	2.93
Max range(km/100)	85.36	4.30	0.21	0.17	85.96	0.22	81.23	4.53	0.20	0.18	90.04	0.22	90.04	0.22
Standard Deviation														
Charge time (hours)	23.60	4.44	0.92	5.42	23.55	0.97	23.29	4.58	0.95	5.59	25.59	0.94	25.59	0.94
Max speed(km/h/100)	46.57	4.33	9.63	7.69	43.71	9.06	44.98	4.73	10.13	6.92	44.30	8.58	44.30	8.58
Max range(km/100)	56.09	4.46	4.18	6.34	57.53	4.79	53.71	4.41	4.27	5.94	59.58	4.95	59.58	4.95
Share	0.29	-6.50			0.27									
Loglikelihood	-2619.79				-2615.88		-2621.79				-2620.19		-2616.68	
Parallel est. time (min)	4.30				15.61		4.68				5.85		16.95	
Sequential est. time (min)	19.66				69.80		21.99				26.66		76.65	
Number of Draws	500				1000		500				500		1000	
Tolerance	-				-		1.00E-03				1.00E-04		1.00E-03	
Scenario	1				2		1(a)				1(b)		2	
Method	MSLE-Analytical						Faster MM							

Table 2.17: Computational Efficiency Ranking

Estimation Methods	**MMNL**		**MON-MNL**		**LML**
	small choice set	**large choice set**	**small choice set**	**large choice set**	**small choice set**
MSLE-Analytical	3	1	1	1	1
MSLE-Numerical	5	4	4	3	3
Faster-MM	1	2	2	2	2
MM	2	3	3	5	5
EM	4	5	5	4	4

Note 1: Parallelization can reduce estimation time of faster-MM and MSLE up to 80% due to storage of information in statistics that are sufficient.

Note 2: Computation efficiency of EM does not improve in parallelization due to communication of large matrices.

Note 3: Faster-MM is as good as MSLE in terms of average percentage bias and finite-sample standard error across all models.

Figure 2.1 shows that the eighth-order LML-polynomial could capture multimodality in the mixing distribution of WTP, which could not be retrieved by the second-order LML-polynomial. Similarly, MON-MNL could also identify two groups with different WTP distributions. Whereas WTP of one of classes MON-MNL has a similar range to LML, the other class has much higher standard deviation and thus a much larger range of WTP measures. For instance, the range of WTP for charge time (per hour) is -7000 to 4000 Rp. in the eighth-order LML-polynomial; the two classes of MON-MNL have ranges of -5,000 to 4,000 Rp. (relatively matching LML) and -20,000 to 10,000 Rp. (much wider than LML), respectively. Whereas the former has mean and median WTP of approximately -1,300 and -1,900 Rp., these estimates for the latter are -2,700 and -900 Rp.

In sum, the performance of the different estimation methods remain intact in the empirical study, but we highlight the sensitivity of the WTP estimates relative to the semiparametric specifications. While the semiparametric modeling literature is growing, better model selection metrics are required to leverage benefits of theses flexible specifications.

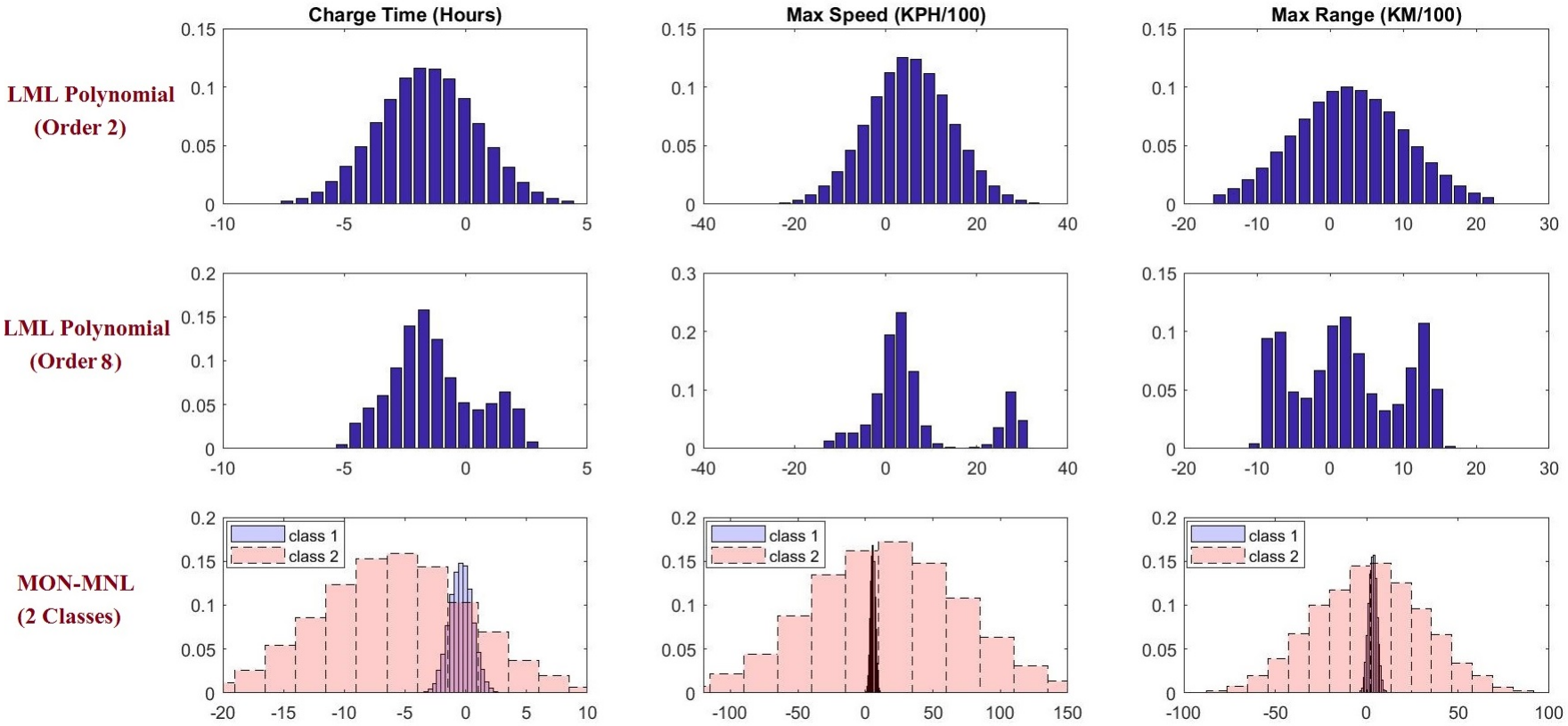

Figure 2.1: Histogram of Willingness to Pay (Rp. thousands)

2.7 Conclusions

The promising computational performance of the minorization-maximization (MM) algorithm in the estimation of parametric mixed multinomial logit (MMNL) models as analyzed by James (2017) encouraged us to derive MM for two semi-paramtric logit models, namely the logit-mixed logit (LML) and the mixture-of-normals multinomial logit (MON-MNL). Whereas the standard EM algorithm (of both LML and MON-MNL) optimizes the loglikelihood of weighted MNL models at each iteration, the simplicity of MM lies in approximating these optimization problems with closed-form parameter updates. In fact, in contrast to EM, we illustrate that MM can leverage advantages of parallel computation. Computational efficiency ranking of the different estimation methods across the different models under analysis is presented and summarized in Table 2.17.

While implementing the algorithm we observed that good behavior of the MM approximation, and thus computational efficiency of the MM algorithm, break down if any of the weighted MNL models has a large choice set. In the LML model, the choice set of the weighted MNL that represents the discrete heterogeneity distribution is intrinsically large (e.g., at least in the order of $J = 500$) as the simulation draws need to adequately cover the parameter space. For situations with a large choice set we have derived a general faster-MM algorithm that simply adjusts the optimization step size using a scalar multiplier.

In separate Monte Carlo studies, we have compared MSLE (with analytical

and numerical expressions of the gradient) against the iterative-optimization estimation algorithms – namely EM, MM, and the proposed faster-MM – to estimate LML and MON-MNL.

For LML, MM lags well behind all algorithms, at the edge of becoming impractical. Whereas the proposed faster-MM reduced MM estimation time by a factor of around 100 and surpassed EM. However, MSLE with an analytical gradient still outperformed faster-MM by a significant margin. In sum, alternative algorithms derived in this study lose their practical appeal for LML estimation.

For the MON-MNL simulated data with small choice sets, we found that MSLE with an analytical gradient was again the fastest estimator, but MM was not far behind. When the number of alternatives increases, MM rapidly deteriorates and even was ourperformed by EM. Thus the main result of James (2017), where MM outperformed MSLE and EM in computational efficiency, appears to be only true for a parametric MMNL model (or one class MON-MNL) with a rather small choice set. The proposed faster-MM was able to reduce MM estimation time by more than 75% and could surpass EM for MON-MNL with a choice set of 50 alternatives (for $J = 50$, 85.4 minutes for faster-MM vs. 53.6 minutes for MSLE-Analytical; cf. 362.5 minutes for MM and 171.3 minutes for EM).

The proposed parallel implementation of faster-MM and MSLE could further reduce estimation time of both by 45% to 80%. Although MSLE with analytical gradient still outperformed faster-MM, parallel computation reduced the scale of the difference between estimation times. Both methods performed equally well

in terms of recovering parameters and estimating asymptotic standard errors. In sum, the (parallel) faster-MM algorithm that we propose and implemented is a competitive alternative to MSLE with analytical gradient for MON-MNL model and is in general better than the most widely used estimator: MSLE with numerical gradients[21].

Finally we argue and illustrate that use of M-step scores in EM and MM can result into wrong estimates of standard errors. The comparable computational performance of faster-MM opens up possibilities of using hybrid faster-MM-MSLE algorithms, which can also streamline the standard errors computation. The hybrid faster-MM-MSLE algorithms are likely to outperform EM-MSLE algorithms in terms of simplicity and computation speed, while maintaining numerical stability.

As a general result, the faster-MM algorithm would be advantageous for models that involve really complex loglikelihood gradients. Furthermore, both the simplicity of coding (only arithmetic operations) and speed of faster-MM make this algorithm an attractive alternative for estimation software that is flexible in terms of structural specifications. Flexible code allows the user to directly write the loglikelihood or the utility functions. Examples of this type of software includes Biogeme (Bierlaire, 2016) and the R package created by the Choice Modelling Centre at Leeds (CMC, 2017), which allow for estimation of any model at

[21] In fact, the comparison of faster-MM against MSLE with numerical gradients, instead of analytical, gradient is more apposite since both methods require similar mathematical rigor and inputs.

the cost of working with numerical gradients. MSLE with analytical gradients may be the fastest option (as it was in the models analyzed in this chapter) but its implementation is not only model- but also specification-specific (current implementations, for example, are reserved to linear-in-parameter utility functions).

We envision as further work testing the performance of the faster-MM algorithm for such complex specifications, for example integrated choice and latent variable (ICLV) models, which are usually very slow to estimate in flexible code that uses numerical gradients. A significant difference in the willingness to pay estimates of the flexible LML and MON-MNL also raises an important questions about model selection from a growing family of semiparametric choice models.

CHAPTER 3

DESIGNED QUADRATURE TO APPROXIMATE INTEGRALS IN MAXIMUM SIMULATED LIKELIHOOD ESTIMATION

3.1 Introduction

Discrete choice models are widely applied across several disciplines such as marketing, economics, and travel behavior. The mixed multinomial logit (MMNL) model currently dominates empirical choice modeling research since it can capture unobserved preference heterogeneity in willingness to pay (WTP) of decision-makers. However, the multinomial probit (MNP) model is also an attractive alternative to specify flexible substitution patterns across alternatives, as well as to jointly model mixed types of dependent variables (Bhat, 2015). In the maximum likelihood estimator of both MMNL and MNP models, choice probabilities involve computation of a multidimensional integral[1] (Train, 2009). In the absence of analytic solutions[2], these integrals are generally approximated through simulation.

In general, the above-mentioned estimation problems include evaluation of

[1] In fact, estimating design criteria in Bayesian D-efficient designs of choice experiments also requires computation of multidimensional integrals (Yu et al., 2010).

[2] Although Bhat (2011) introduced a simulation-free maximum approximate composite marginal likelihood (MACML) estimation approach for MNP model, the Geweke - Hajivassiliou - Keane (GHK) simulator (Geweke et al., 1994) still is more commonly used in practice.

integrals of the following type:

$$\int_{\Gamma} f(\boldsymbol{x})\omega(\boldsymbol{x})d\boldsymbol{x} \approx \sum_{q=1}^{n} f(\boldsymbol{x}_q)w_q,$$

where Γ is a set in the d-dimensional Euclidean space $\mathbb{R}^d$, ω is a probability density function (or positive weight function), and $f(.)$ is generally a conditional likelihood function. Instead of solving the actual integral, simulation-based inference considers a discrete approximation. The objective of computationally efficient simulation is to determine nodes $\boldsymbol{x}_q$ and weights w_q so that integration can be approximated with minimum number of function evaluations (n).

Simulation-based inference in discrete choice models started with Pseudo-Monte Carlo (PMC) methods. As an alternative to PMC, Quasi-Monte Carlo (QMC) methods are now typically used to approximate multidimensional integrals (Bhat, 2001; Train, 2009). More specifically, low-discrepancy sequences[3] such as randomized and scrambled Halton sequences (Bhat, 2003) and modified latin hypercube sampling (MLHS) (Hess et al., 2006) dominate the empirical literature. QMC methods are prefered over PMC because QMC requires fewer draws (i.e., fewer function evaluations) to approximate the integrals due to their excellent coverage properties (Bhat, 2001). Sándor and Train (2004) and Munger et al. (2012) showed superiority of *digital nets* over Halton sequences, but implementation simplicity of the latter makes it a popular alternative in practice.

Empirical unstability of point estimates with a low number of evaluations

[3] Dick and Pillichshammer (2014) illustrates that the lower the discrepancy of a sequence, the smaller will be the error in the Monte Carlo integration.

when using either PMC or QMC has motivated researchers to explore easy-to-implement numerical methods that can accurately approximate the integral of interest with fewer function (i.e., integrand) evaluations (n) than QMC. In this study, we argue and illustrate that recent developments in quadrature methods open such possibilities.

3.1.1 Quadrature Methods and Research Gap

As an alternative to QMC, quadrature has been explored in the discrete choice literature (Heiss and Winschel, 2008; Heiss, 2010; Abay, 2015; Patil et al., 2017; Goos and Mylona, 2018). Quadrature methods mainly differ from QMC in two ways, as quadrature a) generally assumes that the integrand can be approximated on a polynomial space; b) uses deterministic draws (or *nodes*) that carry unequal weights.

The Gaussian quadrature method approximates one-dimensional integrals with just a few nodes.[4] Quadrature can be simply extended to multiple dimensions using the tensor product. However, this multidimensional extension of quadrature suffers from the curse of dimensionality: the number of nodes (i.e., function evaluations) increases exponentially with the number of dimensions, making it impractical beyond 4-5 dimensions. Smolyak (1963) proposed a way

[4] A $K-$times differentiable integrand can be approximated by a polynomial of degree K, and thus the resulting integral with surrogate integrand can be approximated using just $\frac{K+1}{2}$ nodes (Golub and Welsch, 1969).

to extend the univariate quadrature rule to multiple dimensions in a method that is often called sparse grid quadrature (SGQ) in the literature. For example, whereas Gaussian quadrature can exactly compute an integral with a univariate polynomial of order 5 with 3 nodes, the same function in 20 dimensions requires $3^{20} = 3,486,784,401$ and 841 nodes in product rule quadrature and SGQ methods, respectively (Heiss and Winschel, 2008).

Heiss and Winschel (2008) have demonstrated that SGQ performs much better than QMC in estimation of the MMNL model with even up to 20 random parameters. Further, Heiss (2010) combined SGQ with efficient importance sampler (EIS)(Richard and Zhang, 2007) to estimate MNP and panel binary probit models, and demonstrated superiority of this hybrid SGQ-EIS approach over traditional QMC methods.

Even if nodes and weights in SGQ can be pre-computed and stored for reuse as easily as in traditional QMC methods, SQG methods have not been adopted in practice due to three possible reasons. *First*, weights computed in SGQ can be negative. Whereas Heiss and Winschel (2008) discussed this concern as an eventual possibility, the authors claimed not to encounter any such issue – perhaps due to a very simplistic simulation design with a (low-variance) diagonal variance-covariance matrix. In contrast, in our experience we always encountered the issue of negative choice probability estimates for a few individuals coming from negative weights, which in addition to be meaningless numerically led to imaginary (complex) loglikelihood values. Patil et al. (2017) also encountered convergence

issues due to negative weights while applying the SGQ-EIS method in estimation of multinomial probit. *Second*, the required number of nodes to accurately approximate the integral using SGQ depends on the functional properties of the integrand, but the researcher is generally not aware of these properties. *Third*, whereas SGQ reduces the number of nodes significantly as compared to the product rule, the cardinality remains very high relative to that of QMC for high dimensional integrals. Concerns two and three can be illustrated with the following example. If the integrand in a ten-dimensional integral can be well approximated using a 3^{rd}-order polynomial, Gaussian SGQ just needs 21 nodes, but the number of required nodes and thus the number of function evaluations increases to $8,761$ for a 9^{th}-order polynomial (Heiss and Winschel, 2008). The combined consequences of concerns two and three is confirmed by Abay (2015) in estimation of a panel binary probit – SGQ outperforms QMC for dimensions below or equal to 4, but QMC starts dominating SGQ for higher dimensions, and the difference is apparent as panel covariance increases.[5]

3.1.2 Moment-base Quadrature and Contributions

More recent developments in quadrature methods could address the main concerns of SGQ. Whereas Ryu and Boyd (2015) showed that numerical quadrature can be obtained by solving an infinite-dimensional linear program (LP), Jakeman

[5]Note that higher panel covariance in binary probit makes the integrand (i.e., loglikelihood) less smooth leading to a higher order polynomial (i.e., higher number of nodes or function evaluations) are required to approximate the integral at the same level of accuracy.

and Narayan (2018) used the same flexible moment-based optimization framework to obtain a numerical quadrature rule. Very recently, Keshavarzzadeh et al. (2018) simplified this moment-based strategy by solving a relaxed version of the original optimizarion problem and came up with a new numerical quadrature rule known as designed quadrature (DQ).

DQ has many key features. This flexible, new framework allows the researcher to add a constraint to always obtain positive weights. Moreover, DQ rules can be constructed over non-standard geometries of the support of the nonnegative weight function[6] and on more general polynomial spaces (e.g., hyperbolic cross polynomial space) instead of restricting to just total-order polynomial spaces. In fact, DQ requires relatively fewer nodes than SGQ. For example, to approximate a 10 dimensional integral with a polynomial of total order 5 as integrand, DQ requires 148 nodes while nested SGQ needs 201 nodes.[7] To the best of our knowledge, full potential of moment-based numerical quadrature rules have not been explored in the econometrics literature. Thus, the contribution of study is twofold: i) we address the bottlenecks of the traditional SGQ method by applying the re-

[6]For example, Keshavarzzadeh et al. (2018) considered the support of weight function to be "U" shape while generating DQ.

[7]In the absence of information about functional properties of the integrand, the issue of assuming that the pre-specified order of polynomial would approximate the integrand persists in DQ. However, adaptive SGQ methods (Ma and Zabaras, 2009; Brumm and Scheidegger, 2017; Cagnone and Bartolucci, 2017; Bhaduri and Graham-Brady, 2018) are capable of handling this issue. In fact, these adaptive methods are not restricted to polynomial basis functions, and thus hierarchical linear or non-linear basis functions are generally used to capture the local behavior of the integrand. The problem is that these methods are generally computationally expensive and since the basis function is adaptively updated in each dimension based on properties of the integrand, the nodes and weights need to be computed for each problem (i.e., cannot be reused). Cagnone and Bartolucci (2017) show possibilities of parallel computation to make adaptive SGQ faster, but exploring those possibilities is beyond the scope of this study.

cently developed designed quadrature (DQ) method (Keshavarzzadeh et al., 2018) in maximum simulated likelihood estimation of discrete choice models; (ii) using a Monte Carlo study and an empirical application, we show superiority of DQ over QMC methods in estimation of MMNL with varying number of random parameters (3, 5, and 10) and correlation structures (diagonal, low covariance, and high covariance).

The rest of the chapter is organized as follows: section 2 briefly describes the MMNL model and its estimation, section 3 discusses univariate quadrature, multivariate quadrature, and DQ methods, section 4 explains the Monte Carlo simulation design and summarizes corresponding results, section 5 compares QMC methods with DQ on an empirical study, and conclusions and future work are detailed in section 6.

3.2 Mixed Multinomial Logit Model

Consider that the conditional indirect utility derived by decision-maker i from making choice j in choice situation t is:

$$U_{itj} = \boldsymbol{x}_{itj}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i + \varepsilon_{itj}, \tag{3.1}$$

where $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, J\}$, and $t \in \{1, \ldots, T\}$. The covariate vector $\boldsymbol{x}_{itj}$ has a fixed preference parameter vector $\boldsymbol{\alpha}$ and $\boldsymbol{z}_{itj}$ has a random, agent-specific parameter vector $\boldsymbol{\beta}_i$. The preference shock ε_{itj} is independent across individuals, choices

and time, and is identically distributed Type-I Extreme Value. Thus, the probability of choosing alternative j by individual i in choice situation t, conditional on $\boldsymbol{\beta}_i$, has a logit link:

$$P_{itj}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \frac{\exp\left(\boldsymbol{x}_{itj}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itj}^T \boldsymbol{\beta}_i\right)}{\sum_{k=1}^{J} \exp\left(\boldsymbol{x}_{itk}^T \boldsymbol{\alpha} + \boldsymbol{z}_{itk}^T \boldsymbol{\beta}_i\right)}. \tag{3.2}$$

For an individual i who chooses alternative j in choice situation t, we define the indicator $d_{itj} = \mathbb{I}(\text{j chosen}|i,t)$. For the sequence of choices made by individual i, the conditional likelihood $\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_i)$ is:

$$\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_i) = \prod_{t=1}^{T} \prod_{j=1}^{J} [P_{itj}(\boldsymbol{\alpha}, \boldsymbol{\beta}_i)]^{d_{itj}}. \tag{3.3}$$

Consider that the random parameter $\boldsymbol{\beta}_i$ is multivariate normally distributed with mean $\boldsymbol{\gamma}$ and variance-covariance matrix $\boldsymbol{\Delta}$. Thus, the loglilkelihood $\ell(\boldsymbol{\psi})$ of the sample in terms of the unconditional likelihood $P_i(\boldsymbol{\psi})$ of individual i is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{N} \ln\left(P_i(\boldsymbol{\psi})\right) = \sum_{i=1}^{N} \ln\left(\int_{\boldsymbol{\beta}} \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\Delta}) d\boldsymbol{\beta}\right), \tag{3.4}$$

where $\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Delta}\}$.

Since the sample loglikelihood $\ell(\cdot)$ in equation 3.4 is analytically intractable, the parameter vector $\boldsymbol{\psi}$ can be estimated by maximizing the sample's simulated loglikelihood $\widetilde{\ell}(\cdot)$:

$$\widetilde{\ell}(\boldsymbol{\psi}) = \sum_{i=1}^{N} \ln\left(\sum_{r=1}^{R} \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}_{ir}) w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\gamma}, \boldsymbol{\Delta})\right). \tag{3.5}$$

Note that $\boldsymbol{\beta}_{ir}$ and $w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\gamma}, \boldsymbol{\Delta})$ are viewed as nodes and weights in the quadrature method, respectively. In the QMC simulation literature, nodes are generally denoted by draws and the weight $w_i(\boldsymbol{\beta}_{ir}|\boldsymbol{\gamma}, \boldsymbol{\Delta})$ attains the value of $\frac{1}{R}$ for all draws[8].

3.3 Quadrature Methods

3.3.1 Notation

We adopt the notation of Keshavarzzadeh et al. (2018) to illustrate the intuition and key results of different quadrature methods. We reconsider the integral approximation problem:

$$\int_{\Gamma} f(\boldsymbol{x})\omega(\boldsymbol{x})d\boldsymbol{x} \approx \sum_{q=1}^{n} f(\boldsymbol{x}_q)w_q, \tag{3.6}$$

where $\omega(\boldsymbol{x})$ is a given weight function (or a probability density function) whose support is $\Gamma \subset \mathbb{R}^d$. A point $\boldsymbol{x} \in \mathbb{R}^d$ has components $\boldsymbol{x} = \left(x^{(1)}, x^{(2)}, \ldots, x^{(d)}\right)$.

We define $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ as a multi-index, and Λ as a downward closed set[9] of multi-indices:

[8]Note that even though $\boldsymbol{\beta}_{ir}$ is a realization of $\mathcal{N}(\boldsymbol{\gamma}, \boldsymbol{\Delta})$, the model is reparametrized in terms of the Cholesky decomposition of $\boldsymbol{\Delta}$ to ensure positive definiteness. Thus, when approximating the loglikelihood with quadrature or QMC methods, we always work with standard normal distributions.

[9]If $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d$, then $\boldsymbol{\alpha} \leq \boldsymbol{\beta}$ if and only if all component-wise inequalities are true. Using this definition, a multi-index set Λ is called *downward closed* if $\boldsymbol{\alpha} \in \Lambda \implies \boldsymbol{\beta} \in \Lambda \quad \forall\, \boldsymbol{\beta} \leq \boldsymbol{\alpha}$.

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d), \qquad \boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^{d} \left(x^{(j)}\right)^{\alpha_j}, \qquad |\boldsymbol{\alpha}| = \sum_{j=1}^{d} \alpha_j.$$

Our ultimate goal is to construct a set of n points $\left\{\boldsymbol{x}_q\right\}_{q=1}^{n} \subset \Gamma$ and positive weights $w_q > 0$ in equation 3.6, but we attempt to achieve this by enforcing equality in equation 3.6 for a subspace Π of polynomials such that:

$$\begin{aligned} &\int_{\Gamma} f(\boldsymbol{x})\omega(\boldsymbol{x})d\boldsymbol{x} = \sum_{q=1}^{n} f(\boldsymbol{x}_q)w_q, \quad f \in \Pi \\ &\Pi = \operatorname{span}\left\{\boldsymbol{x}^{\boldsymbol{\alpha}} \;\middle|\; \boldsymbol{\alpha} \in \Lambda\right\}. \end{aligned} \tag{3.7}$$

Thus, under the assumption that the integrand $f(\cdot)$ is smooth enough to be approximated in the polynomial subspace Π, solving for $\left\{\boldsymbol{x}_q\right\}_{q=1}^{n}$ and $w_q > 0$ using equation 3.7 should provide a good approximation of the integral in equation 3.6.

Whereas Keshavarzzadeh et al. (2018) proposed a numerical method to solve equation 3.7 for a general polynomial subspaces, we restrict discussion to *total order* (represented by subscript $\mathcal{T}_{(.)}$) polynomial spaces[10] with the total order being r:

$$\Pi_{\mathcal{T}_r} = \operatorname{span}\left\{\boldsymbol{x}^{\boldsymbol{\alpha}} \;\middle|\; \boldsymbol{\alpha} \in \Lambda_{\mathcal{T}_r}\right\}, \quad \text{where } \Lambda_{\mathcal{T}_r} = \left\{\boldsymbol{\alpha} \in \mathbb{N}_0^d \;\middle|\; |\boldsymbol{\alpha}| \leq r\right\}. \tag{3.8}$$

[10] We also tried designed quadrature on *hyperbolic cross* polynomial subspaces, but its performance was poorer than the *total order* polynomial spaces.

3.3.2 Univariate Quadrature

We need to define first the basis for the polynomial space $\Pi_{\mathcal{T}_k}$. Note that a basis of orthonormal polynomials exists with elements $p_m(\cdot)$ such that $\deg p_m = m$. The family of these polynomials satisfies the following recursive relation (Askey, 1975):

$$\begin{aligned} xp_m(x) &= \sqrt{b_m}p_{m-1}(x) + a_m p_m(x) + \sqrt{b_{m+1}}p_{m+1}(x), \\ a_m &= (xp_m, p_m) \quad b_m = \frac{(p_m, p_m)}{(p_{m-1}, p_{m-1})} \end{aligned} \tag{3.9}$$

After characterizing the one-dimension polynomial space, we present a theorem which is the foundation of the Quadrature literature:

Theorem 1 (Gaussian quadrature) Let $x_1, \ldots, x_n$ be the roots of the n^{th} orthogonal polynomial $p_n(x)$ and let $w_1, \ldots, w_n$ be the solution of the system of equations

$$\sum_{q=1}^{n} p_j(x_q)w_q = \begin{cases} \sqrt{b_0}, & \text{if } j = 0 \\ 0, & \text{for } j = 1, \ldots, n-1. \end{cases} \tag{3.10}$$

Then $x_q \in \Gamma$ and $w_q > 0$ for $q = 1, 2, \ldots, n$, and

$$\int_{\Gamma} p(x)\omega(x)dx = \sum_{q=1}^{n} p(x_q)w_q \tag{3.11}$$

holds for all polynomials $p \in \Pi_{\mathcal{T}_{2n-1}}$.

According to Theorem 1, nodes and weights in equation 3.11 (which is a one-dimensional version of equation 3.7) can be exactly obtained by solving the sys-

tem of equations (moment-matching conditions) summarized in equation 3.10. A more intuitive implication of Theorem 1 is that if the integrand can be exactly specified on a polynomial space of order $2n-1$, only n nodes are required to compute the corresponding univariate integral precisely. Golub and Welsch (1969) and Davis and Rabinowitz (2007) provide a detailed procedure to compute this univariate quadrature rule.

3.3.3 Multivariate Quadrature

In product rules, univariate quadrature can be simply extended to multivariate quadrature using a tensor product. More specifically, the weight function $\omega(\boldsymbol{x})$ and its support Γ can be written as follows:

$$\Gamma = \times_{j=1}^{d} \Gamma_j, \qquad \omega(\boldsymbol{x}) = \prod_{j=1}^{d} \omega_j\left(x^{(j)}\right),$$

where $\Gamma_j \subset \mathbb{R}$ is univariate domain and $\omega_j(\cdot)$ is a univariate weight. If $p_n^{(j)}(\cdot)$ is the univariate orthonormal polynomial family corresponding to ω_j over Γ_j, then the family of multivariate polynomials orthonormal under ω can be written as:

$$\pi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{j=1}^{d} p_{\alpha_j}^{(j)}\left(x^{(j)}\right), \qquad \boldsymbol{\alpha} \in \mathbb{N}_0^d.$$

The corresponding polynomial space is: $\Pi_{\mathcal{T}_r} = \text{span}\left\{\pi_{\boldsymbol{\alpha}} \;\middle|\; \boldsymbol{\alpha} \in \Lambda_{\mathcal{T}_r}\right\}$. After characterizing the polynomial subspace, the moment-matching conditions of Theorem 1 can be extended to the multivariate case as follows:

Proposition 1 Let Λ be a multi-index set with $\mathbf{0} \in \Lambda$. Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $w_1, \ldots, w_n$ are the solution of the system of equations

$$\sum_{q=1}^{n} \pi_{\boldsymbol{\alpha}}(\boldsymbol{x}_q) w_q = \begin{cases} 1/\pi_{\mathbf{0}}, & \text{if } \boldsymbol{\alpha} = \mathbf{0} \\ 0, & \text{if } \boldsymbol{\alpha} \in \Lambda \backslash \{\mathbf{0}\} \end{cases} \tag{3.12}$$

then

$$\int_{\Gamma} \omega(\boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x} = \sum_{q=1}^{n} \pi(\boldsymbol{x}_q) w_q \tag{3.13}$$

holds for all polynomials $\pi \in \Pi_{\Lambda}$.

Note that unlike Theorem 1, the above proposition neither guarantees the positivity of weights nor ensures that nodes belong to support Γ. Although sparse grid quadrature (SGQ) provides an efficient way to combine multiple dimensions so as to reduce the function evaluations, it does not provide remedy for these issues. We did not consider SGQ in this study, because: a) in our initial test runs, negative weights in SGQ led to complex (imaginary) loglikelihood values in estimation of MMNL with full variance-covariance matrix; b) based on extensive simulations studies, Keshavarzzadeh et al. (2018) confirmed that designed quadrature (DQ) requires many fewer nodes than SGQ. Heiss and Winschel (2008) can be referred for intuitive and theoretical discussion on SGQ rules.

3.3.4 Designed Quadrature (DQ)

DQ solves a relaxed version ofthe moment-matching conditions given in equation 3.12, which enforces positivity of weights and also ensure nodes to fall in the support of the probability density function. Keshavarzzadeh et al. (2018) reformulates the moment-matching conditions as follows:

For a given index set Λ with size $M = |\Lambda|$, consider the matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ with columns $\boldsymbol{x}_j$, and let $\boldsymbol{w} \in \mathbb{R}^n$ be a vector containing the n weights. Let $\boldsymbol{V}(\boldsymbol{X}) \in \mathbb{R}^{M\times n}$ denote the Vandermonde-like matrix with entries

$$(V)_{k,j} = \pi_{\alpha(k)}\left(\boldsymbol{x}_j\right), \quad k = 1,\dots,M \quad j = 1,\dots,n, \tag{3.14}$$

where elements of Λ are considered with ordering $\boldsymbol{\alpha}(1),\dots\boldsymbol{\alpha}(M)$ and $\boldsymbol{\alpha}(1) = \mathbf{0}$. The system (3.12) can then be written as:

$$\boldsymbol{V}(\boldsymbol{X})\,\boldsymbol{w} = \boldsymbol{e}_1/\pi_{\mathbf{0}}, \tag{3.15}$$

where $\boldsymbol{e}_1 = (1,0,0,\dots,0)^T \in \mathbb{R}^M$. Instead of solving the moment-matching conditions exactly in equation 3.15, Keshavarzzadeh et al. (2018) proposed to obtain the approximate solution $(\boldsymbol{X},\boldsymbol{w})$ that satisfies:

$$\|\boldsymbol{V}(\boldsymbol{X})\,\boldsymbol{w} - \boldsymbol{e}_1/\pi_{\mathbf{0}}\|_2 = \epsilon \geq 0. \tag{3.16}$$

In fact, Keshavarzzadeh et al. (2018) provide bounds on the integral error $\left|\int f(\boldsymbol{x})\omega(\boldsymbol{x})d\boldsymbol{x} - \sum_{q=1}^{n} f(\boldsymbol{x}_q)w_q\right|$ in terms of tolerance ϵ, which is computable for a given quadrature rule.

Thus, for a given polynomial subspace, DQ aims to compute nodes $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \Gamma^n$ and positive weights $\boldsymbol{w} \in (0, \infty)^n$ that solves the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{X}, \boldsymbol{w}} \quad & \|\boldsymbol{V}(\boldsymbol{X})\boldsymbol{w} - \boldsymbol{e}_1/\pi_{\boldsymbol{0}}\|_2 \\ \text{subject to} \quad & \boldsymbol{x}_j \in \Gamma, \qquad j = 1, \ldots, n \\ & w_j > \boldsymbol{0}, \qquad j = 1, \ldots, n. \end{aligned} \tag{3.17}$$

Readers can refer to Keshavarzzadeh et al. (2018) for more insights about strategies (e.g., constrained optimization problem) to solve the above optimization problem.

3.3.5 Discussion

In the context of this study, we explore possibilities of approximating the unconditional choice probability integral (see equation 3.4) in MMNL using DQ. We assume that the conditional choice probability (integrand in equation 3.4) can be approximated on total order polynomial space. Since properties of the integrand vary with the data generating process and are thus not known beforehand, the performance of the approximation will depend on the assumed order of the polynomial space. Moreover, whereas SGQ predetermines the exact number of nodes based on the order (r) of the polynomial space and dimension of the integral, DQ rules can be obtained (i.e., the optimization problem in equation 3.17 can be solved) for various possible number of nodes (n). Thus, for a given integral dimen-

sion, one can generate DQ rules for different total order (r) polynomial spaces and different number of nodes (n).

In both parametric and non-parametric MMNL models, the choice probability integral can generally be reparameterized such that the weight function $\omega(.)$ in DQ turns out to be a probability density function of a standard normal (e.g., normal or lognormal mixing distributions) or a standard uniform distribution (e.g., semi-parametric logit-mixed logit model)[11]. Just as in QMC methods we can generate, store and reuse draws/nodes for both standardized distributions, DQ offers the same flexibility. The researcher can solve the optimization problem in equation 3.17 beforehand for different combinations of dimensions, order of polynomial, and number of nodes, and then reuse the stored nodes and weights.

In sum, DQ may appear more cumbersome than QMC methods at first, but reusability of the nodes and weights not only makes DQ equally easy to implement in practice and in fact, even fewer function evaluations are needed (i.e., lower computation time is achieved). Nevertheless, for a given dimension of integral, whereas QMC needs tuning of the number of draws to get stable parameter estimates, DQ requires to tune the total order of polynomial spaces and the corresponding number of draws. In the next section we conduct a detailed simulation study to make recommendations about selection of these parameters in the context of MMNL.

[11]The support Γ of standard normal and standard distributions are whole real line and $[0, 1]$, respectively

3.4 Monte Carlo Study

3.4.1 Simulation Design

The objective of the simulation study is to evaluate the performance of DQ relative to QMC methods in MMNL estimation. We considered modified latin hypercube sampling (MLHS), which has shown to be superior than other QMC methods such as randomized and scrambled Halton sequences (Hess et al., 2006),[12] as representative of QMC methods. As data generating process (DGP) we considered a sample of 1000 decision makers who are assumed to choose a utility maximizing alternative from a set of five alternatives across five choice situations. Since the number of random parameters governs the dimension of the choice probability integral in MMNL, we compared performance of DQ and MLHS in MMNL with three, five, and ten normally-distributed random parameters. For each random parameter scenario, we considered three covariance structures: zero (diagonal), low, and high covariance across random parameters. The considered low ($\mathbf{\Delta}^5_{low\ cov.}$) and high ($\mathbf{\Delta}^5_{high\ cov.}$) covariance matrices for five random parameters are illustrated below; similar structures were considered for dimensions three and ten. This sensitivity analysis is crucial because the performance of DQ depends on the smoothness of integrand (i.e., conditional likelihood), which in turn depends on the structure of the covariance matrix.

[12] In practice, MLHS and Halton methods are interchangeably used and perform equally well.

$$\mathbf{\Delta}^5_{low\ cov.} = \begin{bmatrix} 1.5 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1.5 \end{bmatrix} \quad \mathbf{\Delta}^5_{high\ cov.} = \begin{bmatrix} 1.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1.5 \end{bmatrix}$$

We generated 250 datasets in total for each covariance structure: 100 datasets for three and five random parameters, and 50 datasets for ten random parameters.[13] For each of 750 datasets, we performed maximum simulated likelihood estimation (with analytical gradient) using a different number of MLHS draws and different total order polynomial subspaces and nodes (or draws) of DQ. We summarize results by computing the following four metrics across resamples: average *loglikelihood* at convergence, finite sample standard error (*FSSE*, or standard deviation of the point estimates), absolute percentage bias (*APB*)[14], and value of the *t-distributed test statistic*[15] under the null hypothesis that the point estimate is equal to the true population parameter. To avoid empirical identification issues, we computed FSSE, APB, and t-value for parameter ratios. We compute these statistics for each parameter, but report averages across all parameters for suc-

[13] We restricted the number of resamples to 50 for ten random parameters due to high computation time.

[14] We compute the absolute percentage bias (APB) of a parameter for a sample as follows: $\text{APB}= \left|\frac{\text{Parameter Estimate - True Parameter Value}}{\text{True Parameter Value}}\right| \times 100$. The mean of APB across all resamples is reported.

[15] We compute the statistic to test the parameter recovery as follows: $\frac{\text{Mean of the Point Estimate across Resamples - True Parameter Value}}{\text{FSSE}}$. As the test statistic gets smaller. we become more confident that the estimated parameter is close to the population parameter.

cinctness. We wrote MATLAB code to generate DQ rules and perform MMNL estimation. DQ rules were generated beforehand, stored, and reused for estimation. We considered tolerance ϵ in equation 3.16 to be 10^{-8}. We performed sensitivity analysis with tighter tolerances but those did not improve accuracy.

3.4.2 Results and Discussion

The results of the Monte Carlo study for random parameters (integral dimensions) three, five, and ten are summarized in Tables 3.1, 3.2, and 3.3, respectively.

As the dimension of the integral increases, the minimum number of nodes required to generate the appropriate DQ rule at a given polynomial order (r, also known as accuracy level) increases. For example, we could generate the DQ rule for higher order $r = 7$ with just 30 nodes for three dimensions (see Table 3.1), but to solve the DQ optimization problem (up to a prespecified tolerance ϵ) for the same order in five dimensions needed more than 100 nodes (see Table 3.2). Also, for a given dimension of the integral, more nodes are required in higher order polynomial spaces. For example, we could generate the DQ rule for ten dimensions with 100 nodes for a polynomial of order $r = 4$, but needed a minimum of 200 nodes for $r = 5$ (see Table 3.3).

We now compare model fit (loglikelihood) of DQ and MLHS. In the diagonal variance-covariance case, DQ outperformed MLHS by a significant margin, even when DQ was generated on polynomial spaces with a relatively low order. For

the five dimensional case, DQ achieved better model fit (loglikelihood: -5355.2) with just 100 nodes at $r = 6$ than 500 MLHS draws (loglikelihood: -5357.4). In fact, DQ with 100 nodes at $r = 4$ (loglikelihood: -5013.6) outperformed MLHS with 500 draws (loglikelihood: -5016.0) in approximating the higher (i.e., ten) dimensional integral.

DQ also outperformed MLHS in the non-diagonal variance-covariance scenarios, but higher order of polynomial subspaces are desirable in this non-independent case. These observations are aligned with intuition: introducing covariance makes the integrand more complex (Abay, 2015), which can be better approximated on higher order polynomial subspaces. For example, in the case of five random parameters with a low covariance DGP, whereas DQ could achieve a model fit of -5794.2 with just 300 nodes at $r = 7$, MLHS required 500 draws to achieve virtually the same model fit; however, 300 nodes of DQ at $r = 5$ were outperformed by 300 MLHS draws. Consistent with intuition, we generally observed that increasing the order of polynomial subspaces results into better model fit. We have seen some exceptions to this trend for three random parameters with non-zero covariance, but for a very low number of draws (ie., 30 and 50) which are often not used in practice (see Table 3.1). As a general trend, across all dimensions and covariance structures, the highest order in DQ (r = 7, 7, and 5 for dimensions 3, 5, and 10) resulted in better model fit than MLHS at a given number of draws.

Table 3.1: Comparison of DQ and MLHS (Monte Carlo, Random Parameters=3)

	(-)Loglikelihood			APB			FSSE			t-value		
Draws	MLHS	DQ		MLHS	DQ		MLHS	DQ		MLHS	DQ	
		r=6	r=7		r=6	r=7		r=6	r=7		r=6	r=7
Diagonal												
30	5752.2	5727.4	5726.2	6.7	5.9	5.8	0.032	0.028	0.028	0.37	0.37	0.35
50	5739.9	5726.2	5725.3	6.1	5.7	5.7	0.030	0.029	0.028	0.35	0.32	0.31
100	5733.1	5725.6	5725.3	5.9	5.7	5.5	0.029	0.028	0.028	0.33	0.32	0.31
150	5730.0			5.8			0.029			0.32		
Low Covariance												
30	5914.8	5851.8	5856.4	26.1	20.4	19.4	0.085	0.068	0.069	0.58	0.45	0.50
50	5856.4	5825.0	5834.2	22.3	17.8	18.3	0.076	0.065	0.069	0.45	0.34	0.43
100	5816.3	5816.0	5805.0	19.6	20.0	18.4	0.071	0.074	0.066	0.35	0.19	0.20
150	5803.6			17.9			0.065			0.25		
High Covariance												
30	5914.7	5863.1	5866.3	14.7	14.0	13.3	0.080	0.071	0.070	0.60	0.61	0.54
50	5862.0	5831.4	5846.8	13.0	10.3	12.1	0.076	0.063	0.072	0.46	0.31	0.38
100	5826.4	5821.4	5815.3	11.5	10.4	10.1	0.070	0.067	0.063	0.35	0.23	0.21
150	5814.6			10.6			0.065			0.26		

Note: APB is absolute percentage bias, FSSE is finite sample standard error, and DQ is designed quadrature.

Table 3.2: Comparison of DQ and MLHS (Monte Carlo, Random Parameters=5)

	(-)Loglikelihood				APB				FSSE				t-value			
Draws	MLHS	DQ			MLHS	DQ			MLHS	DQ			MLHS	DQ		
		r=5	r=6	r=7		r=5	r=6	r=7		r=5	r=6	r=7		r=5	r=6	r=7
Diagonal																
50	5389.4	5361.2			8.7	6.3			0.055	0.033			0.33	0.34		
100	5373.9	5356.6	5355.2		6.7	6.0	5.8		0.037	0.033	0.032		0.26	0.25	0.24	
200	5364.4	5354.8	5354.4	5353.5	6.1	5.9	5.7	5.6	0.034	0.033	0.032	0.032	0.23	0.25	0.22	0.21
300	5360.2	5354.6	5353.6	5353.3	5.8	5.8	5.7	5.6	0.033	0.032	0.032	0.032	0.22	0.25	0.21	0.20
500	5357.4				5.7				0.033				0.22			
Low Covariance																
50	5966.3	5902.8			37.0	32.6			0.097	0.084			0.21	0.57		
100	5882.2	5847.2	5840.8		31.9	29.4	26.5		0.087	0.083	0.076		0.22	0.16	0.29	
200	5829.1	5820.6	5814.2	5811.3	26.7	28.0	26.0	26.2	0.077	0.082	0.074	0.076	0.18	0.13	0.22	0.18
300	5810.0	5817.0	5803.2	5794.2	25.6	28.6	24.6	24.1	0.075	0.080	0.073	0.070	0.18	0.20	0.19	0.17
500	5794.5				23.8				0.071				0.15			
High Covariance																
50	5890.3	5840.9			23.5	20.3			0.100	0.083			0.36	0.60		
100	5818.7	5793.7	5790.9		18.0	18.3	18.0		0.084	0.082	0.082		0.25	0.27	0.31	
200	5772.9	5764.5	5762.1	5760.0	15.4	17.3	15.1	15.7	0.074	0.080	0.071	0.075	0.19	0.19	0.21	0.20
300	5757.1	5763.8	5753.1	5747.3	15.2	16.6	14.4	14.1	0.073	0.077	0.069	0.069	0.18	0.21	0.20	0.17
500	5743.6				13.8				0.068				0.16			

Note: APB is absolute percentage bias, FSSE is finite sample standard error, and DQ is designed quadrature.

Table 3.3: Comparison of DQ and MLHS (Monte Carlo, Random Parameters=10)

	(-)Loglikelihood			APB			FSSE			t-value		
Draws	MLHS	DQ		MLHS	DQ		MLHS	DQ		MLHS	DQ	
		r=4	r=5		r=4	r=5		r=4	r=5		r=4	r=5
Diagonal												
50	5044.4			26.7			0.112			0.72		
100	5032.1	5013.6		22.5	15.2		0.109	0.085		0.59	0.44	
200	5022.1	5013.2	5010.8	17.3	12.3	11.0	0.097	0.068	0.067	0.47	0.43	0.33
300	5018.1	5011.0	5010.8	14.1	11.1	11.1	0.085	0.064	0.069	0.37	0.39	0.33
500	5016.0	5011.0		11.9	9.1		0.074	0.052		0.33	0.32	
Low Covariance												
50	6182.4			125.7			0.232			0.34		
100	6126.2	6078.6		104.2	85.0		0.198	0.160		0.27	0.31	
200	6075.8	6050.0	6042.7	80.0	66.1	68.0	0.159	0.124	0.130	0.21	0.25	0.25
300	6054.4	6031.6	6032.8	67.0	64.7	64.5	0.126	0.125	0.123	0.20	0.24	0.23
500	6027.4	6013.8		60.0	56.2		0.118	0.108		0.18	0.20	
High Covariance												
50	5943.0			86.6			0.205			0.46		
100	5893.0	5863.3		83.8	70.9		0.199	0.174		0.35	0.35	
200	5860.4	5844.2	5842.5	69.5	59.6	58.4	0.174	0.153	0.150	0.28	0.26	0.21
300	5842.1	5834.2	5832.9	63.5	60.7	54.7	0.161	0.158	0.139	0.23	0.28	0.21
500	5822.2	5818.6		47.4	45.1		0.127	0.116		0.17	0.19	

Note: APB is absolute percentage bias, FSSE is finite sample standard error, and DQ is designed quadrature.

As expected, across DQ and MLHS all parameter recovery metrics – APB, FSSE, and t-value – decrease with an increase in the number of draws (or nodes). Consistent with model fit, DQ surpassed MLHS by a significant margin in recovering true parameters if the variance-covariance matrix is diagonal. For instance, DQ could achieve lower values of all parameter recovery metrics with 200 nodes (on polynomial space of order $r = 7$) than those of 500 MLHS draws for the DGP with five random parameters (see Table 3.2). In fact, DQ also performed better than MLHS across correlated covariance structures, but at higher order polynomial subspaces. For example, in DGP with five highly correlated random parameters, APB, FSSE, and t-value using 300 MLHS draws are 15.2%, 0.073, and 0.18 respectively, but for 300 DQ nodes whereas at $r = 5$ these values are relatively higher – 16.6%, .077, and 0.21, they are relatively lower – 14.1%, 0.069, and 0.17 at $r = 7$ (see Table 3.2).

In sum, better model fit and more precise parameter recovery of DQ across all dimensions and covariance structures make DQ a strong substitute to QMC methods in practice.

3.5 Empirical Study

We now compare the performance of DQ and QMC while studying the preference of travelers in New York City (NYC) for mobility-on-demand (MoD) services (e.g., Uber and UberPool).

3.5.1 Experiment Design

We conducted a stated preference survey in NYC. The survey included a discrete choice experiment (DCE) in which each respondent was asked to choose the best and the worst travel mode from a set of three choices: Uber (without ridesharing), UberPool (with ridesharing), and their current travel mode (the one used most often on their most frequent trips). We first conducted a pilot study (N=298) using D-efficient design with zero priors in February 2017. We then used prior parameter estimates from the pilot study to create a pivot-efficient design[16] with 6 blocks (7 choice situations per block). Table 3.4 shows the attribute levels of the DCE design and an instance of choice situation. More details about the experiment design can be found in Liu et al. (2018). We conducted the main study during October-November 2017. After data validation tests, preferences of 1507 (out of 1689) respondents were used in estimation.

[16] In pivot-efficient designs, attribute levels shown to the respondents are pivoted from reference alternatives for each respondent. In this study, the travel mode used on the most frequent trips was considered as the reference alternative.

Table 3.4: Experiment Design for Mode Choice Study

Attribute Levels in the Experiment Design			
	Uber (Without Ridesharing)	**UberPool (With Ridesharing)**	**Current Mode**
Walking and Waiting Time	25%, 50%, 75%, 100%	25%, 50%, 75%, 100%	asked (100%)
In-vehicle Travel Time	80%, 95%, 110%, 125%	90%, 105%, 120%,135%	asked (100%)
Trip Cost Per Mile ($) (Excluding Parking Cost)	0.55, 0.70, 0.85, 1.0, 1.2	0.45, 0.60, 0.70, 0.80	asked or computed
Parking Cost	0	0	asked
Powertrain	Gas, Electric	Gas, Electric	Gas
Automation	Yes, No	Yes, No	No
Instance of a Choice Situation			
	Uber (Without Ridesharing)	**UberPool (With Ridesharing)**	**Current Mode: Car**
Walking and Waiting time	6 minutes	9 minutes	12 minutes
In-vehicle Travel Time	38 minutes	50 minutes	48 minutes
Trip Cost (Excluding Parking Cost)	$11	$8	$6
Parking Cost	–	–	$6
Powertrain	Electric	Gas	Gas
Automation	Service with Driver	Automated (No Driver)	–

Note: All % are relative to the reference alternative.

3.5.2 Estimation and Results

We considered marginal utilities of all five alternative-specific variables to be normally-distributed. This specification led to a five dimensional integral in MMNL estimation. We also considered randomized and scrambled Halton draws along with MLHS and DQ. The number of draws/nodes was varied from 50 to 500 and total order (r) of polynomial subspaces in DQ ranged from 5 to 7. We considered 50 different starting values and for each starting value, 19 models were estimated considering different QMC draws and DQ nodes.

Table 3.5: Comparison of -Loglikelihood Values in the Case Study

Number of Draws	MLHS	Halton Draws	Designed Quadrature		
			r=5	r=6	r=7
50	8206.5	8245.4	8233.5		
100	8168.5	8182.9	8149.5	8151.7	
200	8142.1	8151.1	8155.2	8144.5	8124.2
300	8134.1	8142.1	8155.2	8129.8	8121.7
500	8128.1	8135.2			

Table 3.5 summarizes the average of model fit across different starting values. In this study, MLHS draws resulted in better model fit than Halton draws across all considered scenarios. The performance of DQ is consistent with the Monte Carlo study – whereas QMC methods dominated DQ at lower order $r = 5$, DQ generated at higher orders 6 and 7 always outperformed QMC methods across all considered draws. In fact, 200 nodes in DQ at order $r = 7$ could achieve better model fit (-8124.2) than those of 500 Halton (-8135.2) or MLHS (-8128.1) draws.

Table 3.6: Comparison of Estimates and Standard Errors in the Case Study

	Estimates					z-scores				
	MLHS		DQ			MLHS		DQ		
Draws	**200**	**500**	**300 (r=6)**	**200 (r=7)**	**300 (r=7)**	**200**	**500**	**300 (r=6)**	**200 (r=7)**	**300 (r=7)**
Mean										
OVTT/100 (min)	-1.65	-1.67	-1.98	-1.66	-1.84	-3.89	-3.68	-4.47	-4.45	-4.48
IVTT/100 (min)	-10.76	-10.94	-11.18	-11.36	-10.85	-17.96	-17.82	-18.16	-16.74	-17.74
Trip Cost/10 ($)	-3.32	-3.41	-3.45	-3.79	-3.62	-19.20	-18.81	-19.02	-18.63	-19.08
Electric?	-0.38	-0.38	-0.39	-0.41	-0.41	-6.15	-6.08	-6.08	-6.35	-6.30
Automation?	-0.49	-0.50	-0.53	-0.53	-0.54	-7.48	-7.40	-7.49	-7.33	-7.59
Cholesky components										
L11	0.10	0.79	0.33	1.04	2.16	0.17	1.25	0.54	1.60	3.20
L21	0.58	0.14	1.04	1.45	1.09	0.66	0.15	0.95	1.65	1.27
L22	9.09	8.72	7.03	8.66	8.95	13.35	12.83	10.37	10.91	13.79
L31	0.20	0.09	0.49	1.26	0.70	1.24	0.47	3.28	8.19	3.73
L32	1.33	1.28	1.46	1.86	1.30	8.84	7.38	9.89	12.88	8.71
L33	2.68	2.21	2.10	2.97	2.30	14.47	10.80	10.76	15.96	12.69
L41	0.01	-0.09	0.02	-0.05	-0.22	0.07	-1.02	0.21	-0.56	-2.44
L42	0.14	0.13	0.24	0.23	0.22	1.32	1.21	2.77	2.37	2.22
L43	0.34	0.24	0.25	0.36	0.30	3.44	2.30	2.40	3.60	3.12
L44	0.22	0.28	0.69	0.02	0.68	1.64	2.22	5.87	0.18	6.08
L51	-0.01	-0.10	-0.02	-0.10	-0.23	-0.13	-0.92	-0.16	-1.01	-2.31
L52	0.06	0.06	0.22	0.17	0.13	0.51	0.50	2.28	1.70	1.19
L53	0.32	0.21	0.25	0.32	0.33	2.92	1.80	2.21	2.98	3.09
L54	0.22	0.22	0.56	0.14	0.39	1.28	1.27	3.84	1.05	2.67
L55	-0.16	-0.19	0.14	-0.55	-0.27	-0.92	-1.31	1.12	-3.72	-1.76
Loglikelihood	-8142.1	-8128.1	-8129.8	-8124.2	-8121.7					

Table 3.6 shows the parameter estimates and z-scores for selected MLHS draws and DQ nodes. The mean estimates of MLHS and DQ are similar and in fact, z-score values are also in a similar range. The Cholesky components (e.g., L22 and L33) which are statistically significant in MLHS remains significant in DQ and as expected, corresponding point estimates are also more stable across the considered draws. A few Cholesky components (e.g., $L11$ and $L31$) which are not statistically significant in MLHS with 500 draws, appear significant in DQ with 300 draws at order $r = 7$. However, this observation requires further validation in other case studies.

3.6 Conclusions

In this study, we have proposed the use of designed quadrature (DQ) to approximate multi-dimensional integrals in maximum simulated likelihood estimation of discrete choice models. We have compared performance of DQ with traditionally used QMC methods in a Monte Carlo and an empirical case study.

Whereas traditional sparse grid quadrature methods suffer from the problem of complex-valued loglikelihood due to negative weights, DQ could estimate MMNL smoothly for DGPs with varying covariance structures, thanks to positivity of weights. The simulation and empirical study confirmed that DQ requires fewer function evaluations than QMC if the variance-covariance matrix is diagonal. In DGPs with non-diagonal matrices and varying covariance structures, DQ

always outperforms MLHS in terms of model fit and parameter recovery when the quadrature rule is generated on higher order polynomial subspaces.

In sum, features like positivity of weights, computational efficiency due to fewer function evaluations, and easy implementation due to reusability of quadrature rules make DQ a potentially attractive alternative to QMC methods. As a future work, we plan to test its sensitivity relative to sample size, number of choice situations, number of alternatives, and other discrete choice models (e.g., multinomial probit, and semi-parametric logit models).

Furthermore, to ensure better performance of DQ over QMC, the key question is: for a given dimension and number of draws, on what maximum order of polynomial subspaces, DQ rule can be generated? Thus, also as future work, taking advantage of the re-usability feature of DQ we plan to create software that can store the DQ rules on the highest possible order for commonly encountered dimensions, weight functions, and the number of nodes. With said software, DQ is as easy to use as any other QMC method, but with better performance. In other words, similar to QMC methods, the user would just need to choose the number of draws for the given dimension and software can provide the best DQ rule.

CHAPTER 4

A CONTINUOUS-MULTINOMIAL RESPONSE MODEL WITH A T-DISTRIBUTED ERROR KERNEL

4.1 Introduction

Discrete-response (e.g., count, ordered, binary, multinomial) models are popular across various disciplines such as applied economics, transportation, marketing, and political science. In these models, error structures are specified at different modeling levels because the researcher does not have full information about the data generating process (DGP). For instance, the total indirect utility in additive random utility maximization (ARUM) (McFadden, 1973) based choice models is specified as the sum of a deterministic index (depending on observables) and a random error term or taste shock. Such error specifications in empirical studies are generally governed by ease of estimation rather than structural appropriateness (Vijverberg and Vijverberg, 2016). In particular, without worrying about the characteristics of the data in hand, Gumbel/extreme-value (logit) or normal (probit) error kernels are commonly used. These traditional logit and probit links have been replaced by a t-distributed error kernel, i.e. a robit link (Liu, 2004), in multilevel modelling applications to handle fat-tailed error distributions. Moreover, a t-distribution with an estimable degree-of-freedom (DOF) actually generalizes logit and probit.[1] However, we are not aware of the use of the robit link in model-

[1]The t-distribution with about seven and *a large* (above thirty) DOF approximates logistic and normal distributions, respectively.

ing multinomial responses, perhaps because the benefits of the generalization are unclear and the estimation of the resulting model is cumbersome.

In this study, we present the first application of a multinomial response model with a t-distributed error kernel, multinomial robit (MNR) model henceforth. Whereas the proposed MNR model retains all merits of the multinomial probit (MNP) model[2], we illustrate three additional advantages of adopting MNR in practice. First, we highlight that the robit link is superior to the probit link in estimating and predicting preferences in unbalanced datasets where one or more alternatives have small shares. Second, we parameterize the DOF of the t-distributed error kernel as a function of demographics and show how this specification can help in capturing decision uncertainty of decision-makers that standard compensatory ARUM models cannot account for. Third, if the error distribution in the true DGP is fat-tailed, unlike probit, the robit link can retrieve the true model parameters at excellent accuracy. We numerically establish this benefit of MNR over MNP in a Monte Carlo study.

Given the growing interest in the joint modeling of mixed datasets across multiple disciplines (see De Leon and Chough, 2013, for applications), we further extend MNR to a generalized continuous-multinomial (GCM) response model with a t-distributed error kernel that facilitates simultaneous consideration of multiple multinomial and multiple continuous dependent variables. We note that the ad-

[2]Similar to MNP, MNR allows for flexible substitution patterns – correlations across indirect utilities of alternatives – without necessarily including variation in parameters across decision-makers.

vantages of considering a t-distributed (over a normally-distributed) error kernel are even more evident in the GCM response model because fat-tailed distributions are more commonly observed in continuous outcomes. Moreover, the joint GCM model offers several advantages, namely: a) statistically efficient estimation, b) easier hypothesis testing and better power of statistical tests, c) avoidance of inconsistencies in a situation when continuous and multinomial endogenous outcomes affect each other (see Bhat et al., 2015, for a detailed discussion). For example, the GCM response model is appropriate for the joint modeling of commute distance (continuous variable) and choice of residential location (multinominal variable) because both outcomes clearly affect each other. Joint modeling is thus desirable, but it is generally intractable due to an absence of convenient distributions to represent the conditional and/or joint relationship between the outcomes. The proposed GCM response model with a t-distributed error kernel (henceforth, GCM-t) could not have been estimated conveniently without an elegant statistical property of the t-distribution: the conditional distribution of a joint/multivariate t-distribution is also a t-distribution (Ding, 2016).

The contribution of this study is thus threefold. First, we illustrate the importance of t-distributed error kernels in multinomial choice modeling. Second, we derive a full-information maximum likelihood procedure to estimate MNR and GCM-t models. The likelihood expressions of both models involve evaluation of high-dimensional multi-variate-t-cumulative density (MVTCD) functions. We adopt the composite marginal likelihood (or paired-likelihood) approach to first decompose the multi-dimensional MVTNCD integral into multiple pairs and thus

reduce the integral dimensionality (see Varin et al., 2011; Xu and Reid, 2011). Subsequently, similar to the Geweke-Hajivassiliou-Keane (GHK) method to simulate multivariate normal cumulative density functions (Genz, 1992; Hajivassiliou et al., 1996), we use a separation-of-variables approach to simulate MVTCD functions (Genz and Bretz, 1999). Third, we numerically verify the statistical properties of the maximum likelihood estimator of the GCM-t model in a Monte Carlo study. We also compare the performance of GCM-t and GCM with normally-distributed error kernel (GCM-N) models in a second simulation study. Finally, we validate the simulation results and highlight the advantages of the robit link (over probit) in an empirical study in the context of policies for on-street parking with charging facilities for electric vehicles. The empirical data looks into the adoption of electric vehicles and the outcomes of interest are a household's vehicle miles traveled (continuous) and vehicle-purchase preferences (multinomial).

The remainder of this chapter is organized as follows. We present the contextual literature review in section 4.2; section 4.3 details specification of the GCM-t model and derives its estimator; section 4.4 illustrates advantages of adopting choice models with t-distributed error kernels in practice; 4.5 presents a comprehensive Monte Carlo study and highlights the benefits of GCM-t over GCM-N; section 4.6 validates the simulation-based findings in the empirical study; and, finally, section 4.7 concludes and discusses avenues of future research.

4.2 Literature Review

Regarding flexibility in modeling limited dependent variables, there is an extensive literature covering semi-parametric error distributions for binary- and multinomial-response models. However, we restrict our discussion to flexible, parsimonious parametric error specifications, and highlight research gaps that this study addresses.

To overcome constraints from the assumption of symmetric and thin-tailed error kernels in logit and probit models, existing research offers several alternative error specifications with one or two additional shape parameters. In statistical terms, whereas an error kernel with one additional parameter (e.g., t-distribution) allows for the trade-off between skewness and kurtosis, two additional parameters (e.g., skew t-distribution) accommodate several choices of both skewness and kurtosis.

The scobit model derived from a Burr-10 distribution (Nagler, 1994), the robit model which considers a t-distribution with estimable DOF (Liu, 2004), and the skew-probit model which assumes a skew-normal distribution with estimable skew parameter (Bazán et al., 2010) are among well-known binary response models that use error kernels with one additional parameter. Error kernels with two additional parameters, namely skewed t-distribution (Kim et al., 2007) and the generalized Tukey lambda (GTL) family of distributions (Vijverberg and Vijverberg, 2016) have also been applied for binary outcomes.

Some of these binary-response models with a flexible error structure have been extended to the multinomial dependent variables. Castillo et al. (2008) and Fosgerau and Bierlaire (2009) derived a multinomial choice model considering a Weibull distribution on the error term in a specification that is also known as the weibit model. Li (2011) proposed a generalized method to construct asymmetric multinomial choice models for a family of error distributions with heteroskedastic variance, which also nests the weibit and logit models. Recently, Nakayama and Chikaraishi (2015) derived a unified multinomial choice model using the q-generalized-extreme-value (q-GEV) distribution with an estimable shape parameter and applied their model to transportation network assignment problems. Brathwaite and Walker (2018) identified that all these flexible multinomial models impose restrictions on the magnitude and/or sign of the index function. To address this concern, Brathwaite and Walker proposed a generalized link function that eliminates the need for such restrictions.

Note that all the multinomial choice models overviewed above have tractable closed-form choice probability expressions but suffer from a major limitation – the joint modeling of multiple types of dependent variables (e.g., continuous, ordinal, and count) and the inclusion of spatial and social dependencies in these models are computationally intractable, if not impossible, due to an increase in the dimensionality of integration (see Guevara et al., 2009, for a discussion on the curse of dimensionality in choice models).

Error kernels with skew-normal, t, or skew-t distributions have appropriate

statistical properties for such joint structural modeling, but the use of these flexible distributions in multinomial choice models is cumbersome due to open-form choice probability expressions as well as noted inference issues.[3]

The skew-normal distribution has been used at a few instances in the mixed MNP model but not to specify the error kernel, rather to model unobserved preference heterogeneity (random parameter choice models as in Bhat and Sidharthan, 2012). Since random taste variations and error term heterogeneity are confounded (Brownstone and Train, 1998), the mixed MNP specification can be viewed as an indirect utility with a non-random index function and a skew-normal error term. Bhat et al. (2015) and Bhat et al. (2017) have also used the skew-normal distribution to account for non-normality in latent constructs within an integrated choice and latent variable model (Ben-Akiva et al., 2002) and in the error kernel of an ordered response model, respectively.

However, we are not aware of the application of t- or skew-t-distributed error kernels in multinomial choice models, which are commonly used in binary, linear mixed or multilevel, and censored linear regressions (Pinheiro et al., 2001; Liu, 2004; Koenker and Yoon, 2009; Marchenko and Genton, 2012; Wang et al., 2018) due to their ability to model fat-tailed distributions. This study particularly contributes to the literature by first illustrating the statistical and behavioral implications of using a fat-tailed error kernel in multinomial-response models and

[3]Models with skew-normal and skew-t error kernel can encounter inference problems due to eventual singularity of the Fisher information matrix (when direct parameterization is used) and violation of asymptotic theory for centered parameterization (see Pewsey, 2000; Azzalini and Arellano-Valle, 2013, for a discussion on these issues).

proposing the first multinomial choice model with the t-distributed error kernel.[4]

4.3 Methodology

In this section, we first discuss separate models for multiple-continuous and multiple-discrete responses with t-distributed error kernels. We then combine these models to derive a generalized continuous-multinomial response model with a t-distributed error kernel (GCM-t) and outline steps for implementation of its full-information maximum likelihood estimator.[5] In the proposed model formulation, we do not consider network, social, or spatial effects and therefore the utility of an individual is independent of other individuals in the sample. Thus, without loss of generality, we derive the model and estimation procedure for a single individual.

4.3.1 Continuous variable model

Consider a standard linear regression setup: $y_h = \boldsymbol{\gamma_h}^\intercal \boldsymbol{X}_h + \xi_h$, where h is the index of a continuous outcome $h = \{1, 2, \ldots, H\}$, y_h and ξ_h are the corresponding dependent variable and a t-distributed error term with DOF δ, and $\boldsymbol{\gamma}_h$ and $\boldsymbol{X}_h$ are respectively

[4]The skew-t kernel could additionally account for asymmetry of the error distribution, but we do not consider it because inference in such models would have similar issues as we encounter in models with skew-normal kernel (Azzalini and Arellano-Valle, 2013).

[5]Since multinomial robit (MNR) model is a special case of the GCM-t model, we do not explicitly discuss its estimator.

$(s \times 1)$ vectors of coefficients and exogenous regressors.

Rewriting the regression equation in matrix form leads to: $\boldsymbol{y} = \text{diag}(\boldsymbol{\gamma X}^{\top}) + \boldsymbol{\xi}$, where $(\boldsymbol{y})_{H\times 1} = [y_1, y_2, \ldots, y_H]^{\top}$ is a vector of continuous outcomes and $(\boldsymbol{\xi})_{H\times 1} = [\xi_1, \xi_2, \ldots, \xi_H]^{\top}$ is a t-distributed error vector with DOF δ, and $(\boldsymbol{\gamma})_{H\times s} = [\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \ldots, \boldsymbol{\gamma_H}]^{\top}$ and $(\boldsymbol{X})_{H\times s} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_H}]^{\top}$ are matrices of coefficients and exogenous variables. Note that $\boldsymbol{y} \sim \text{MVT}_H\left[\text{diag}(\boldsymbol{\gamma X}^{\top}), \boldsymbol{\Xi}, \delta\right]$ where $\boldsymbol{\Xi}$ is the variance-covariance matrix of $\boldsymbol{\xi}$.

We consider the same DOF across all elements of the error vector because the exact distribution of linear or non-linear combinations of two t-distributed random variables with arbitrary DOF values is not known (Ahsanullah et al., 2014). Jones (2002) allows marginal distributions to have an arbitrary DOF in the case of a bivariate t-distributed random variable, but its extension to the multivariate case is not straightforward.

4.3.2 Choice model

Let i be the index for a nominal outcome $i \in \{1, 2, \ldots, I\}$, and k be the index of alternatives in each nominal outcome $k \in \{1, 2, \ldots, i_K\}$. Then, we can write the indirect utility of alternative k in the i^{th} nominal variable as $U_{ik} = \boldsymbol{\beta}_{ik}^{\top}\boldsymbol{z}_{ik} + \epsilon_{ik}$, where $\boldsymbol{z}_{ik}$ and $\boldsymbol{\beta}_{ik}$ are $(g \times 1)$ vectors of exogenous variables and coefficients, and ϵ_{ik} is a t-distributed error term with DOF δ.

If we define the total number of alternatives $I_K = \sum_{i=1}^{I} i_K$, the indirect utility vector $(\boldsymbol{U})_{I_K \times 1} = [\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_I]^\intercal$ where $\boldsymbol{U}_i = [U_{i1}, U_{i2}, \ldots, U_{ii_K}]$, the coefficient matrix $(\boldsymbol{\beta})_{I_K \times g} = [\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}, \ldots, \boldsymbol{\beta}_{11_K}, \ldots, \boldsymbol{\beta}_{II_K}]^\top$, the exogenuous variable matrix $(\boldsymbol{z})_{I_K \times g} = [\boldsymbol{z}_{11}, \boldsymbol{z}_{12}, \ldots, \boldsymbol{z}_{11_K}, \ldots, \boldsymbol{z}_{II_K}]^\top$, and the error vector $(\boldsymbol{\epsilon})_{I_K \times 1} = [\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_I]^\intercal$ where $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{ii_K}]$, we can write the distribution of the indirect utility as $\boldsymbol{U} \sim \text{MVT}_{(I_K \times I_K)} \left[\text{diag}(\boldsymbol{\beta}\boldsymbol{z}^\intercal), \boldsymbol{\Lambda}, \delta\right]$.

Since only differences in utility matter, the difference of error terms $(\overline{\boldsymbol{\epsilon}})$ is identifiable after fixing the scale of utility. In fact, we normalize the top diagonal element of the covariance matrix of error differences to 1 to fix scale of utility. We create a *transformation matrix* $(\boldsymbol{D})$ to convert the normalized variance-covariance of error differences $(\overline{\boldsymbol{\Lambda}}_{(I_K-1)\times(I_K-1)})$ into the undifferenced error variance-covariance matrix $(\boldsymbol{\Lambda}_{I_K \times I_K})$ using $\boldsymbol{\Lambda} = \boldsymbol{D}\overline{\boldsymbol{\Lambda}}\boldsymbol{D}^\intercal$. We provide details of creating the transformation matrix $(\boldsymbol{D})$ and an illustration of this operatoor in appendix C.1.1 (cf. Bhat and Sidharthan, 2012). The indirect utility can thus be written as $\boldsymbol{U} \sim \text{MVT}_{I_K} \left[\text{diag}(\boldsymbol{\beta}\boldsymbol{z}^\intercal), \boldsymbol{D}\overline{\boldsymbol{\Lambda}}\boldsymbol{D}^\intercal, \delta\right]$.[6]

4.3.3 Joint Model Specification

In order to write a joint model of the continuous and nominal variable, we define $\boldsymbol{YU} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{U} \end{bmatrix}$. Thus, the distribution of $\boldsymbol{YU} \sim \text{MVT}_{H+I_K}[\boldsymbol{B}, \boldsymbol{\Sigma}, \delta]$ where

[6]The matrix $\overline{\boldsymbol{\Lambda}}$ can be block-diagonal and still the dependencies across alternatives and nominal variables are parsimoniously generated by a single DOF parameter.

$\boldsymbol{B} = \begin{bmatrix} \text{diag}(\boldsymbol{\gamma X}^\top) \\ \text{diag}(\boldsymbol{\beta z}^\top) \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Xi} & \text{Cov}(\boldsymbol{\xi}, \boldsymbol{\epsilon}) \\ \text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\xi}) & \boldsymbol{\Lambda} \end{bmatrix}$. If $\overline{\boldsymbol{\Sigma}}$ is the normalized (up to scale) covariance matrix of the joint differenced error $\begin{bmatrix} \boldsymbol{\xi} \\ \overline{\boldsymbol{\epsilon}} \end{bmatrix}$, which is identified, the undifferenced full variance-covariance matrix ($\boldsymbol{\Sigma}$) can be obtained from $\overline{\boldsymbol{\Sigma}}$ using the modified transformation matrix $\boldsymbol{D}_m$ as follows: $\boldsymbol{\Sigma} = \boldsymbol{D}_m \overline{\boldsymbol{\Sigma}} \boldsymbol{D}_m^\top$. Appendix C.1.2 provides the details of creating $\boldsymbol{D}_m$, together with an example.

4.3.4 Joint Model Estimation

Similar to MNP estimation, we work with utility differences using the chosen alternative as base. To perform this operation, we construct the *utility difference operator* $\boldsymbol{M}$ of size $(H + I_K - I) \times (H + I_K)$ using the algorithm given in appendix C.1.3. We transform the original mean and the variance-covariance matrix using $\boldsymbol{M}$, and thus derive the distribution of the joint variable $\widetilde{\boldsymbol{YU}}$ in utility-differences ($\overline{\boldsymbol{U}}$) space. We obtain $\widetilde{\boldsymbol{YU}} \sim \text{MVT}_{H+I_K-I}\left(\widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{\Sigma}}, \delta\right)$, where $\widetilde{\boldsymbol{B}} = \boldsymbol{MB}$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{M\Sigma M}^\top$.

Consider the partition of $\widetilde{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{\Sigma}}$ into the continuous and choice (discrete) model (in utility differences) as follows: $\widetilde{\boldsymbol{B}} = \begin{bmatrix} \widetilde{\boldsymbol{B}}_y \\ \widetilde{\boldsymbol{B}}_{\overline{U}} \end{bmatrix}_{(H+I_K-I)\times 1}$ and $\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \widetilde{\boldsymbol{\Sigma}}_y & \widetilde{\boldsymbol{\Sigma}}_{y,\overline{U}} \\ \widetilde{\boldsymbol{\Sigma}}_{\overline{U},y} & \widetilde{\boldsymbol{\Sigma}}_{\overline{U}} \end{bmatrix}_{(H+I_K-I)\times(H+I_K-I)}$.

The conditional distribution of the utility difference vector is also t-distributed

(Ding, 2016):

$$\overline{\boldsymbol{U}}|\boldsymbol{y} \sim \text{MVT}_{I_K-I}\left(\overleftrightarrow{\boldsymbol{B}}_{\overline{U}}, \overleftrightarrow{\boldsymbol{\Sigma}}_{\overline{U}}, \overleftrightarrow{\delta}\right), \tag{4.1}$$

where

$$\begin{aligned}
\overleftrightarrow{\boldsymbol{B}}_{\overline{U}} &= \widetilde{\boldsymbol{B}}_{\overline{U}} + \widetilde{\boldsymbol{\Sigma}}^{\intercal}_{\boldsymbol{y},\overline{\boldsymbol{U}}}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\right)^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{B}}_{\boldsymbol{y}}) \\
\overleftrightarrow{\boldsymbol{\Sigma}}_{\overline{U}} &= \left[\frac{\delta+\alpha}{\delta+H}\right]\left(\widetilde{\boldsymbol{\Sigma}}_{\overline{U}} - \widetilde{\boldsymbol{\Sigma}}^{\intercal}_{\boldsymbol{y},\overline{\boldsymbol{U}}}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\right)^{-1}\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y},\overline{\boldsymbol{U}}}\right) \\
\overleftrightarrow{\delta} &= \delta + H \\
\alpha &= (\boldsymbol{y} - \widetilde{\boldsymbol{B}}_{\boldsymbol{y}})^{\intercal}\left(\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\right)^{-1}(\boldsymbol{y} - \widetilde{\boldsymbol{B}}_{\boldsymbol{y}})
\end{aligned}$$

Thus, the joint likelihood can be written as :

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) =& \Pr(\boldsymbol{y})\Pr(1_k = 1_m, 2_k = 2_m, \ldots, I_k = I_m|\boldsymbol{y}) \\
=& f_H\left(\boldsymbol{y}|\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}, \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}, \delta\right)\left[\int_{-\infty}^{\overleftrightarrow{\boldsymbol{B}}_{\overline{U}}} f_{(I_K-I)}\left(r|\overleftrightarrow{\boldsymbol{B}}_{\overline{U}}, \overleftrightarrow{\boldsymbol{\Sigma}}_{\overline{U}}, \overleftrightarrow{\delta}\right)d\boldsymbol{r}\right]
\end{aligned} \tag{4.2}$$

where i_m is the chosen alternative corresponding to the i^{th} nominal variable, f_H is the probability density function of the H-variate t-distribution, and $\boldsymbol{\theta}$ is a vector of identified parameters $\{\boldsymbol{\gamma}, \boldsymbol{\beta}, \text{vec}(\overline{\boldsymbol{\Sigma}})\}$.[7]

In order to ensure positive definiteness of the normalized covariance matrix of error difference ($\overline{\boldsymbol{\Sigma}}$), we work with its Cholesky decomposition in estimation. We

[7]Note that $\overline{\boldsymbol{\Sigma}}$, the normalized variance-covariance matrix of the joint differenced error, is identified and other matrices ($\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$) are derived from it using $\boldsymbol{D}_m$ and $\boldsymbol{M}$ matrices. Moreover, vec($\overline{\boldsymbol{\Sigma}}$) vectorizes the unique element of a matrix $\overline{\boldsymbol{\Sigma}}$.

effectively consider the normalized Cholesky factorization such that the top diagonal element of the error differenced covariance matrix of every nominal variable ($\overline{\mathbf{\Lambda}}_i$) is fixed to 1. Details of normalization can be found in appendix C.1.4.

The full-information maximum likelihood estimator requires maximization of the joint likelihood function. All numerical maximization routines require to evaluate the likelihood function at a given parameter vector, which we illustrate in Algorithm 4. This computation involves evaluation of a $H-$dimensional multivariate-t-probability-density (MVTPD) function and a $(I_K - I)-$dimensional multivariate-t-cumulative-density (MVTCD) function. The MVTCD function does not have a closed-form expression and thus requires the use of simulation-aided inference. We use a separation-of-variables (SoV) approach to compute the MVTCD function (Genz and Bretz, 1999), which is detailed in section 4.3.4. Similar to other simulation-based function evaluation procedures, the SOV approach also suffers from the *course of dimensionality* coming from the increase in integral dimensionality. As a result, simulation-based evaluation of the function not only loses accuracy, but computation time also becomes unmanageable (Bhat, 2003; Craig, 2008). To reduce the dimension of this integration, we adopt the composite marginal likelihood (CML) method, which we briefly discuss in section 4.3.4.

Algorithm 4: An algorithm to compute the loglikelihood of GCM-t model

Input data: $\{\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, H, I, N\}$, where N is the number of decision-makers;
Input parameters: $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\beta}, \text{vec}(\boldsymbol{L}_{\overline{\boldsymbol{\Sigma}}})\}$, where $\boldsymbol{L}_{\overline{\boldsymbol{\Sigma}}}$ is the lower triangular Cholesky matrix of $\overline{\boldsymbol{\Sigma}}$;
Step 1: Reparametrize $\boldsymbol{L}_{\overline{\boldsymbol{\Sigma}}}$ using the procedure given in appendix C.1.4 and compute $\overline{\boldsymbol{\Sigma}}$;
Step 2: Create *modified transformation matrix* $\boldsymbol{D}_m$ using the procedure provided in appendix C.1.2;
Step 3: Compute undifferenced error variance-covariance matrix: $\boldsymbol{\Sigma} = \boldsymbol{D}_m \overline{\boldsymbol{\Sigma}} \boldsymbol{D}_m^{\intercal}$;
Step 4: Compute mean of the joint dependent variable $\boldsymbol{YU}$ for the sample: $\boldsymbol{B} = \begin{bmatrix} \text{diag}(\boldsymbol{\gamma}\boldsymbol{X}^{\intercal}) \\ \text{diag}(\boldsymbol{\beta}\boldsymbol{z}^{\intercal}) \end{bmatrix}$;
Step 5:
for *(n in 1 to N)* **do**
Construct utility difference generator $\boldsymbol{M}$ using the algorithm 6 in appendix C.1.3;
Compute mean $\widetilde{\boldsymbol{B}} = \boldsymbol{M}\boldsymbol{B}$ and covariance matrix $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^{\intercal}$ in the utility difference space;
Obtain the conditional distribution of the utility difference $(\overline{\boldsymbol{U}}|\boldsymbol{y})$ using equation 4.1;
Compute the likelihood for the decision-maker n using equation 4.2:

- Compute PDF for the continuous variable: $\mathcal{L}^n_{cont} = f_H\left(\boldsymbol{y}|\widetilde{\boldsymbol{B}}_{\boldsymbol{y}}, \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}, \delta\right)$;
- Use CML approach (section 4.3.4) to decompose the CDF: $\mathcal{L}^n_{nom} = \int_{-\infty}^{\overleftrightarrow{\boldsymbol{B}}_{\overline{U}}} f_{(I_K - I)}\left(r|\overleftrightarrow{\boldsymbol{B}}_{\overline{U}}, \overleftrightarrow{\boldsymbol{\Sigma}}_{\overline{U}}, \overleftrightarrow{\delta}\right) d\boldsymbol{r}$;
- Compute the decomposed CDF using MVTCD function simulator (section 4.3.4);

end
Step 6: Compute the sample loglikelihood: $\mathcal{LL} = \sum_{n=1}^{N} \log\left[\mathcal{L}^n_{cont} \cdot \mathcal{L}^n_{nom}\right]$;

Separation-of-Variables Approach to evaluate MVTCD Function

The underlined concept of this simulator is similar to the GHK simulator for the MVNCD function (Genz, 1992; Hajivassiliou et al., 1996). In the SOV approach, a $p-$dimensional integral of the MVTCD function is decomposed into $(p-1)$ unidimensional integrals using the Cholesky decomposition of the covariance matrix. These independent unidimensional integrals are sequentially evaluated based on the realization of all previous integrals. The approach presented below is adopted from Genz and Bretz (1999).

We present this approach for computing the area under the curve[8] of a probability density function of the p-dimensional t-distributed random variable ($\boldsymbol{r} \sim \text{MVT}_p(\boldsymbol{0}, \boldsymbol{\Omega}, \delta)$)[9] between $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = \int_{\boldsymbol{a}}^{\boldsymbol{b}} f(\boldsymbol{r}|\boldsymbol{\Omega}, \delta) d\boldsymbol{r} = \frac{\Gamma\left(\frac{\delta+p}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)(\pi\delta)^{\frac{p}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}}} \int_{\boldsymbol{a}}^{\boldsymbol{b}} \left(1 + \frac{\boldsymbol{r}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{r}}{\delta}\right)^{-\frac{(\delta+p)}{2}} d\boldsymbol{r} \tag{4.3}$$

Next, $\boldsymbol{\Omega} = \boldsymbol{L}_{\boldsymbol{\Omega}} \boldsymbol{L}_{\boldsymbol{\Omega}}^\top$, where $\boldsymbol{L}_{\boldsymbol{\Omega}}$ is the lower triangular Cholesky factor. Then, by change of variable, $\boldsymbol{r} = \boldsymbol{L}_{\boldsymbol{\Omega}} \boldsymbol{w}$ and $\boldsymbol{r}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{r} = \boldsymbol{w}^\top \boldsymbol{w}$, $d\boldsymbol{r} = |\boldsymbol{L}_{\boldsymbol{\Omega}}| d\boldsymbol{w}$, $d\boldsymbol{r} = |\boldsymbol{\Omega}|^{\frac{1}{2}}$. Let

[8] The area under the probability density function between $\boldsymbol{a}$ and $\boldsymbol{b}$ is the same as the CDF evaluated at point $\boldsymbol{b}$ when $\boldsymbol{a}$ is $-\infty$.

[9] We set location parameter to $\boldsymbol{0}$ without loss of generality because if $\boldsymbol{g} \sim \text{MVT}_p(\boldsymbol{\mu}, \boldsymbol{\Omega}, \delta)$ and $\boldsymbol{r} \sim \text{MVT}_p(\boldsymbol{0}, \boldsymbol{\Omega}, \delta)$, then $\boldsymbol{g} = \boldsymbol{r} + \boldsymbol{\mu}$.

$\kappa_\delta^p = \frac{\Gamma\left(\frac{\delta+p}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)(\pi\delta)^{\frac{p}{2}}}$ such that equation 4.3 can be rewritten as:

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = \kappa_\delta^p \int\limits_{\boldsymbol{a} \le \boldsymbol{L_\Omega w} \le \boldsymbol{b}} \left(1 + \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\delta}\right)^{-\frac{(\delta+p)}{2}} d\boldsymbol{w} \tag{4.4}$$

where $\boldsymbol{a} \le \boldsymbol{L_\Omega w} \le \boldsymbol{b}$ is the same as $\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \le \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ L_{p1} & L_{p2} & L_{p3} & L_{pp} \end{bmatrix} \le \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$

Let $\overline{a}_i = \frac{\left(a_i - \sum_{j=1}^{i-1} L_{ij} w_i\right)}{L_{ii}}$ and $\overline{b}_i = \frac{\left(b_i - \sum_{j=1}^{i-1} L_{ij} w_i\right)}{L_{ii}}$. Also note that

$$\left(1 + \frac{\boldsymbol{w}^\top \boldsymbol{w}}{\delta}\right) = \left(1 + \frac{w_1^2}{\delta}\right)\left(1 + \frac{w_2^2}{\delta + w_1^2}\right) \ldots \left(1 + \frac{w_p^2}{\delta + \sum_{j=1}^{p-1} w_j^2}\right) \tag{4.5}$$

Thus, equation 4.4 can be rewritten as :

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = \kappa_\delta^p \int_{\overline{a}_1}^{\overline{b}_1} \left(1 + \frac{w_1^2}{\delta}\right)^{-\frac{(\delta+p)}{2}} \int_{\overline{a}_2}^{\overline{b}_2} \left(1 + \frac{w_2^2}{\delta + w_1^2}\right)^{-\frac{(\delta+p)}{2}} \cdots \int_{\overline{a}_p}^{\overline{b}_p} \left(1 + \frac{w_p^2}{\delta + \sum_{j=1}^{p-1} w_j^2}\right)^{-\frac{(\delta+p)}{2}} d\boldsymbol{w} \tag{4.6}$$

Consider $w_i = u_i \sqrt{\frac{\delta + \sum_{j=1}^{i-1} w_j^2}{\delta + i - 1}}$. We can rewrite equation 4.6 by substituting w_i as

follows:

$$T_p(\boldsymbol{a},\boldsymbol{b},\boldsymbol{\Omega},\delta) = \kappa_\delta^p \sqrt{\left(\frac{\delta}{\delta+1}\frac{\delta}{\delta+2}\cdots\frac{\delta}{\delta+p-1}\right)}\cdots \int_{\hat{a}_1}^{\hat{b}_1}\left(1+\frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} \int_{\hat{a}_2}^{\hat{b}_2}\left(1+\frac{u_2^2}{\delta+2-1}\right)^{-\frac{(\delta+2)}{2}} \cdots \int_{\hat{a}_p}^{\hat{b}_p}\left(1+\frac{u_p^2}{\delta+p-1}\right)^{-\frac{(\delta+p)}{2}} d\boldsymbol{u} \tag{4.7}$$

where $\hat{a}_i = \overline{a}_i \sqrt{\frac{\delta+i-1}{\delta+\sum_{j=1}^{i-1} w_j^2}}$ and $\hat{b}_i = \overline{b}_i \sqrt{\frac{\delta+i-1}{\delta+\sum_{j=1}^{i-1} w_j^2}}$. Further, equation 4.7 can be rewritten as

$$T_p(\boldsymbol{a},\boldsymbol{b},\boldsymbol{\Omega},\delta) = \left[\kappa_{\delta+1-1}^1 \int_{\hat{a}_1}^{\hat{b}_1}\left(1+\frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} du_1\right]\left[\kappa_{\delta+2-1}^1 \int_{\hat{a}_2}^{\hat{b}_2}\left(1+\frac{u_2^2}{\delta+2-1}\right)^{-\frac{(\delta+2)}{2}} du_1\right] \cdots \left[\kappa_{\delta+p-1}^1 \int_{\hat{a}_p}^{\hat{b}_p}\left(1+\frac{u_p^2}{\delta+p-1}\right)^{-\frac{(\delta+p)}{2}} du_p\right] \tag{4.8}$$

The derivation of equation 4.8 from equation 4.6 is illustrated for a bivariate t-cumulative distribution function ($p = 2$) in appendix C.2. Next, we substitute $u_i = t_{\delta+i-1}^{-1}(z_i)$ where $t_{\delta+i-1}(z_i) = \kappa_{\delta+i-1}^1 \int_{-\infty}^{z_i}\left(1+\frac{s^2}{\delta+i-1}\right)^{-\frac{\delta+i}{2}} ds$ is the cumulative density function (CDF) of the univeriate t-distribution with DOF $\delta + i - 1$, and thus $dz_i = k_{\delta+i-1}^1\left[1+\frac{u_i^2}{\delta+i-1}\right]^{-\frac{\delta+i}{2}} du_i$.

Finally, equation 4.8 can thus be rewritten as:

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = \int_{d_1}^{e_1} \int_{d_2}^{e_2} \int_{d_3}^{e_3} \cdots \int_{d_p}^{e_p} dz_1 dz_2 dz_3 \ldots dz_p \tag{4.9}$$

where d_i and e_i are CDF of t-distributed random variable with DOF $\delta + i - 1$ at points $\hat{a}_i$ and $\hat{b}_i$, respectively. After the change of variables $z_i = d_i + \phi_i(e_i - d_i)$, equation 4.9 becomes:

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = (e_1 - d_1) \int_0^1 (e_2 - d_2) \cdots \int_0^1 (e_p - d_p) \int_0^1 d\boldsymbol{\phi} = \underbrace{\int_0^1 \int_0^1 \int_0^1 \cdots \int_0^1}_{(p-1)} f(\boldsymbol{\phi}) d\boldsymbol{\phi} \tag{4.10}$$

which is an integral of $f(\boldsymbol{\phi}) = \prod_{i=1}^{p} (e_i - d_i)$ over the $(p-1)$-dimensional unit hypercube. This integral can be evaluated using different Quasi-Monte-Carlo or randomized lattice rule methods.

Composite Marginal Likelihood Approach

Computation of the joint likelihood function in equation 4.2 requires the evaluation of a high-dimensional integral. The dimension of this integral grows with the number of alternatives per nominal variable and also with the number of nominal variables. For example, if there are ten nominal variables, each with six alternatives, the integral would be fifty-dimensional.

Rather than directly evaluating such high-dimensional integrals, we use the CML approach (also known as paired-likelihood approach) for simplification. The

CML method breaks down the joint likelihood function into multiple pairs, decreasing the dimension of the integral. More specifically, if choices made by an individual across all nominal variables is an event, this event is represented as pairwise observations in CML:

$$\mathcal{L}_{CML}(\boldsymbol{\theta}) = f_H\left(\boldsymbol{y}|\widetilde{\boldsymbol{B}}_y, \widetilde{\boldsymbol{\Sigma}}_y, \delta\right)\left(\prod_{i=1}^{I-1}\prod_{j=i+1}^{I} \Pr(i_k = i_m, j_k = j_m|\boldsymbol{y})\right) \tag{4.11}$$

where i_m represents the chosen alternative for the i^{th} nominal variable. In the CML expression above, the first term is the same as the MVTPD function, but the second term corresponds to the pairing between nominal variables with the highest dimension of integration being equal to $2[max(i_K \forall i)]$. Thus, by employing the CML approach, the dimension of integration in the above example can be reduced to ten from fifty. Of course, the CML approach is only applicable if there are three or more nominal variables.

A comprehensive discussion on the CML approach is outside the scope of this chapter. Readers can refer to Varin and Vidoni (2005) and Varin et al. (2011) for a detailed discussion on the CML approach and Bhat (2014) for its derivation and application in the context of discrete choice models. Apart from well-established asymptotic properties (Bhat, 2014), Bhat and co-authors have tested finite sample properties of CML by applying it to complex econometric models and have observed satisfactory results (Bhat and Sidharthan, 2012; Bhat et al., 2015, 2017).

4.4 Implications of using GCM-t in practice

4.4.1 Class imbalance

Class imbalance, a very high market share of few alternatives relative to others, is often encountered in choice modeling applications such as residential location choice, travel mode choice, and credit card ownership. We take an example of binary-response data to illustrate the importance of using the robit link in datasets with class imbalance. Consider a scenario where a commuter chooses an alternative between car (C) and bicycle (B). We use a data generating process with a much higher share of car than bicycle, which is reflected in the index functions: $V_C = 0.5HI - 0.2S$ and $V_B = -1 - 0.9HI + 0.5S$, where HI and S are indicators for high income and student commuters, respectively.

In these class imbalance situations, accurate prediction of choices is challenging. A good model should ideally predict a higher probability of choosing bicycle (the alternative with a very low market share) when a commuter actually chooses bicycle. We compare predicted choice probabilities for all four demographic groups under normally-distributed and t-distributed error kernels with varying DOFs (Table 4.1). The predicted probabilities of choosing bicycle by a high-income non-student commuter using probit and robit (with 0.1 DOF) links are 0.01 and 0.38, respectively. These values are 0.38 and 0.46 for a low-income student commuter. Clearly, improvement in the predicted probability of bicycle

using robit is higher when the difference between the index function values of the alternatives is higher.[10]

In sum, the flat-tailed nature of the t-distributed kernel is able to better predict the probability of choosing bicycle for all demographic groups, but the extent of the improvement over probit is higher if the difference in index values lies more toward the tails of the error distribution. Thus, t-distributed error kernels increase the likelihood of predicting correct travel mode assignment in such imbalanced datasets.

4.4.2 Behavioral implications

Intuitively, decision rules can be viewed as a mapping from attributes of alternatives to observed choices, which is governed by a latent construct. The latent construct includes an index function (or deterministic utility) and an idiosyncratic error term. Adoption of different decision rules manifests into different latent constructs and, thus, different mappings. Most of the previous studies have modeled different decision rules by modifying the index function. In fact, it turns out that changing the linear index function to non-linear functions in the latent construct of the fully-compensatory RUM-based model can translate it into a model with non-compensatory decision rules (Swait, 2001; Elrod et al., 2004; Martínez et al.,

[10]The difference between the index function values of alternatives is the highest for a high-income non-student commuter and is the lowest for a low-income student commuter.

2009).[11]

We argue that a variation in DOF of the t-distributed error kernel across decision-makers provides a flexible way to model "decision uncertainty" – a degree of (un)certainty that decision-makers hold in their choices relative to the variation in the indirect utility of any alternative – without modifying the index function in the latent construct. We illustrate this choice behavior using a plot of the choice probability vs. the utility at different DOF in a binary choice scenario (see Figure 4.1). Since the relationship between the choice probability and utility is steeper for the respondents with a higher DOF, choice of these decision-makers is more sensitive to a variation in utility, but in a narrower range. More specifically, *decision uncertainty* of the respondents with the DOF of 10 is much larger as compared to those with the DOF of 0.1.

In discrete choice experiments, *decision uncertainty* can depend on the familiarity and experience of the decision-maker about the situation encountered during the experiment, among many other factors. Decision-makers belonging to a specific demographic group can be just more certain about their choices than others. Such choice behaviors cannot be modeled using standard logit and probit links, even when accounting for preference heterogeneity. Also, neglecting *decision uncertainty* results into underestimation of welfare measures (Dekker et al., 2016). A few studies have quantified the *decision uncertainty* by asking follow-up questions

[11]Whereas the decision-maker is assumed to trade all attributes of alternatives and to choose an alternative with the maximum indirect utility in the compensatory RUM framework, non-compensatory rules allow the decision-maker to choose or reject an alternative based on the value of even a single attribute (Schoemaker, 2013).

after each choice task and incorporating these self-reported responses as explanatory variables or in other structural forms (Lundhede et al., 2009; Olsen et al., 2011; Beck et al., 2013; Börger, 2016).

What we propose is to capture individual-specific *decision uncertainty* by parameterizing the DOF of the error kernel as a function of characeteristics of the decision maker. Since the DOF of each individual is obtained as a byproduct of estimation, GCM-t implicitly captures decision-uncertainty behavior without imposing additional econometric structure and also reduces the cognitive burden of respondents by avoiding the need of asking additional questions.

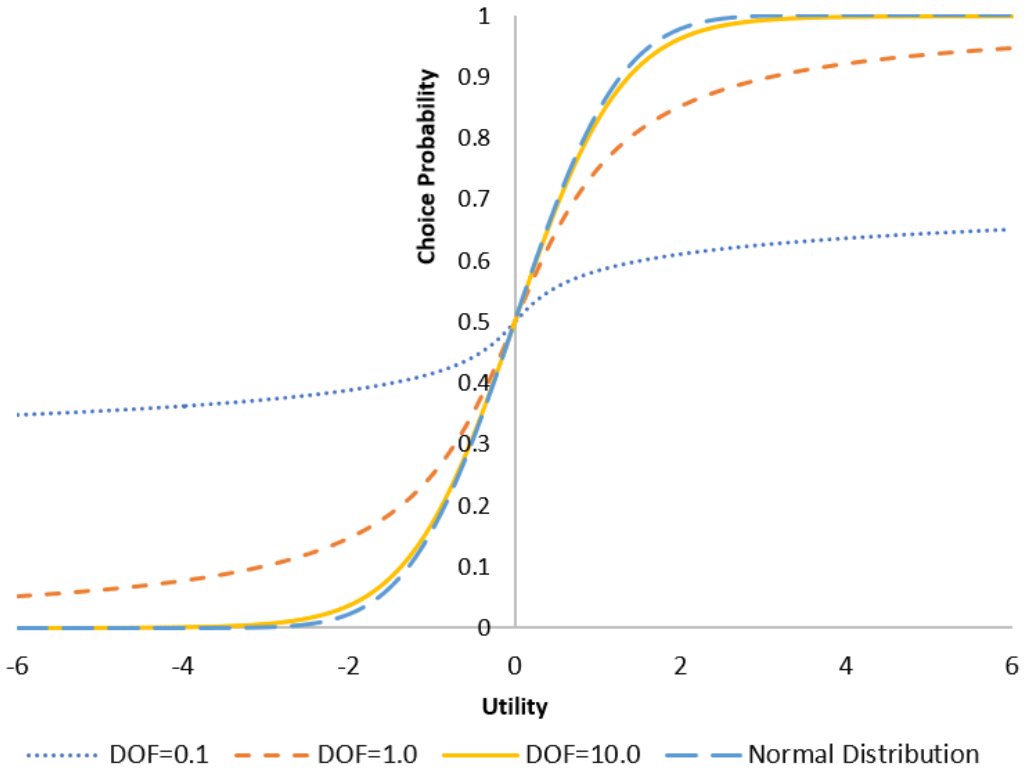

Figure 4.1: Cumulative density function of t- and normally-distributed random variables.

Table 4.1: Class-imbalance example under Probit and t-distributed error kernel

Demographic profile	t-distributed kernel						Probit kernel	
	DOF = 0.1		DOF = 0.5		DOF = 1.0			
	P(car)	P(bicycle)	P(car)	P(bicycle)	P(car)	P(bicycle)	P(car)	P(bicycle)
High-income and non-student	0.62	0.38	0.80	0.20	0.88	0.12	0.99	0.01
High-income and student	0.60	0.40	0.76	0.24	0.83	0.17	0.96	0.04
Low-income and non-student	0.58	0.42	0.70	0.30	0.75	0.25	0.84	0.16
Low-income and student	0.54	0.46	0.58	0.42	0.59	0.41	0.62	0.38

4.5 Monte Carlo study and results

Since this is the first study to use a t-distributed error kernel in multinomial response models, we conduct a simulation study to numerically assess the statistical properties (e.g., recovery of true parameters) of the maximum likelihood estimator of the GCM-t model. In another simulation study, we compare the performances of GCM-t and GCM-N models under thick- and thin-tailed error distributions and highlight the consequences of misspecified error kernels.

The context of our simulation study is joint modeling of the commute distance and residential location choice, i.e. integrated land-use transportation modeling. For residential location, we consider a nominal variable with five population-density-based alternatives: 0 – 99, 100-499, 500 – 1499, 1500 – 1999 and 2000 or more households/square mile. The considered joint model is shown in equation 4.12.

To generate all three exogenous indicators, we take a draw from a standard uniform distribution. If the sampled value is higher than 0.5, the indicator takes a value 1; otherwise, it takes a value 0. That is, for a certain household if the sampled value is 0.64, 0.32, and 0.75, then the generated household is a high-income household with no children, but with a bachelor's degree holder. All the assumed parameter values and their directions are intuitive and consistent with the literature. For example, a high-income household with children prefers to live in a low-density area (Paleti et al., 2013). Similarly, assuming that the commute dis-

tance precedes the choice of residential location (Clark et al., 2003; Rashidi et al., 2012), the likelihood of living in high-density areas decreases with the increase in the commute distance.

Equation 4.13 shows the considered covariance matrix ($\overline{\boldsymbol{\Sigma}}$) of the joint-differenced error, which is normalized up to scale. After taking a draw from this covariance matrix, response variables are generated. Since the commute distance has a structural relationship with the choice of residential location, it is first obtained using the continuous variable model and is then used as an explanatory variable in the choice model to determine the residential location of a household.

$$\boldsymbol{YU} = \begin{bmatrix} \boldsymbol{y} \\ \hline \boldsymbol{U} \end{bmatrix} = \begin{bmatrix} \text{Commute distance} \\ \hline \text{Density 2000+} \\ \text{Density 0-99} \\ \text{Density 100-499} \\ \text{Density 500-1499} \\ \text{Density 1500-2000} \end{bmatrix}$$

$$= \begin{bmatrix} 1.00 & 0.50 & 0.75 & -0.50 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -1.50 & 1.00 & 0.90 & 0.00 & 1.00 \\ -1.30 & 0.90 & 0.80 & 0.00 & 0.90 \\ -1.20 & 0.80 & 0.70 & 0.00 & 0.80 \\ -1.00 & 0.70 & 0.60 & 0.00 & 0.70 \end{bmatrix} \begin{bmatrix} \text{Constant} \\ \text{High-income household} \\ \text{Household with children} \\ \text{Household with a bachelor's degree holder} \\ \text{Commute distance} \end{bmatrix} \tag{4.12}$$

$$\overline{\Sigma} = \left[\begin{array}{c|cccc} \mathbf{1.50} & \mathbf{0.30} & \mathbf{0.40} & \mathbf{0.60} & \mathbf{0.50} \\ \hline \mathbf{0.30} & 1.00 & 0.50 & 0.50 & 0.50 \\ \mathbf{0.40} & 0.50 & \mathbf{1.10} & 0.50 & 0.50 \\ \mathbf{0.60} & 0.50 & 0.50 & \mathbf{1.20} & 0.50 \\ \mathbf{0.50} & 0.50 & 0.50 & 0.50 & \mathbf{1.30} \end{array}\right] \tag{4.13}$$

We initially estimated all elements of the covariance matrix, but such flexible specification resulted in convergence issues – a few elements of the matrix converged to values near zero, leading to a non-positive-definite covariance matrix. This empirical identification concern is commonly encountered, forcing researchers to adopt a diagonal error-covariance matrix. Such concerns are generally not reported, but we are aware of only handful of studies with a non-diagonal error-covariance structure in multinomial choice models with open-form expressions of choice probabilities (e.g., Paleti et al., 2013; Bhat, 2015).

In the simulation study, we estimate diagonal elements of the error-covariance matrix of the choice model and the elements representing the correlation between continuous and multinomial parts of the joint model. The estimated elements of the covariance matrix are in bold in equation 4.13. We note that the diagonal error-covariance in a choice model with a t-distributed error kernel does not translate into independence across alternatives as long as the common DOF is finite.

4.5.1 Statistical properties of GCM-t estimator

To test the statistical properties of the GCM-t model, we consider the above DGP with DOF(δ) 1 (DGP-I) and 12 (DGP-II). Whereas a DOF value of 1 represents a distribution far from normal with flat probability curves and long thick tails, a value of 12 mimics the normal error kernel with steep probability curves and thin tails. In both DGPs, we take a sufficiently large sample of 3000 individuals to circumvent the effect of sample size on parameter recovery. Further, we generate 150 resamples for each DGP to ensure that statistical properties are not affected by the choice of the number of repetitions. To evaluate the MVTCD function, we use 200 Halton draws in the earlier discussed separation-of-variables approach. We have also tested sensitivity of parameter estimates and standard errors by increasing the number of draws from 200 to 500 for a few resamples, but have not observed any improvement in results.[12] We compute the following performance measures across resamples:

Mean estimated value (MEV): Average value of the estimated parameter across all resamples.

Mean absolute bias (MAB): Average bias (|true value-MEV|) of the parameter estimates.

Absolute percentage bias (APB): |MAB/true value|×100.

[12]200 Halton draws have proven to be sufficient to estimate up to eight-dimensional integrals in the GHK simulator for the MNP estimation (Patil et al., 2017).

Finite sample standard error (FSSE): Standard deviation of the parameter estimates across all resamples.

Asymptotic standard error (ASE): Average of standard error values across all resamples, which are obtained using the sandwich estimator. For the sufficient estimator, FSSE and ASE values are close to each other.

Coverage probability (CP): Proportion of times the 95% confidence interval contains the true value.

Power of the test: Proportion of times the null hypothesis is rejected (i.e., |t-statistic| is greater than 1.96).

A lower APB, a ratio of ASE and FSSE closer to 1, higher CP, and higher power are desirable for a better statistical performance of the estimator (Koehler et al., 2009).

The resulting performance measures for DGP-I and DGP-II are summarized in Tables 4.2 and 4.3, respectively. We first highlight important insights related to parameter recovery (i.e., bias). The mean APB values of the parameter vector (γ) in the continuous variable model are 1.54% and 0.46% for DGP-I and DGP-II, respectively. These results not only indicate excellent recovery of γ vector, but also suggest improvement in its recovery with the increase in the DOF. However, parameters (β) associated with the nominal response model appear to be rather difficult to recover accurately irrespective of the DOF. The mean APB values of the β vector are 14.79% and 16.90% for DGP-I and DGP-II, respectively. These values for the covariance matrix ($\overline{\Sigma}$) are 14.16% and 11.37%, respectively, suggesting that the recovery of the error-covariance matrix is better for the DGP with a higher

DOF value. The bias in parameter estimates is substantially higher than the ones obtained when the specification is kept the same but the normally-distributed error replaces the t-distributed error kernel in the DGP and in the estimation. More specifically, the mean APB values of $\boldsymbol{\beta}$ and $\overline{\boldsymbol{\Sigma}}$ in the corresponding GCM-N model are 4.88% and 7.30%, respectively.[13] This comparison suggests that allowing for flexibility in the parametric distribution of the error kernel beyond mean and variance parameters might lead to a deterioration in the recovery of model parameters. Bhat and Sidharthan (2012) also observed a higher bias in the parameter estimates of the mixed multinomial choice model with a skew-normal distribution than those of the corresponding mixed MNP model. The recovery of the DOF parameter (δ) in both cases is good, with APB values of 8.24% and 12.37%, respectively.

We now discuss performance measures related to model inference – ratio of asymptotic standard error and finite sample standard error (ASE/FSSE), coverage probability (CP), and power. ASE/FSEE values of all parameters $\{\boldsymbol{\gamma}, \boldsymbol{\beta}, \overline{\boldsymbol{\Sigma}}, \delta\}$ are close to 1 for both DGPs, suggesting that the proposed estimator is sufficient. However, consistent with the findings of the parameter recovery, mean values of these ratios are much closer to 1 for DGP-II $\{0.99, 0.98, 0.95, 0.98\}$ as compared to those of the DGP-I $\{1.06, 0.81, 0.76, 1.10\}$.

[13] We have also conducted a Monte Carlo study for the corresponding GCM-N model. We used the GHK simulator with 200 Halton to evaluate the MVNCD function. The detailed results are available upon request.

Whereas the mean CP values of the γ vector are close to 0.95 for both DGPs, these values are much lower for other parameters. Particularly, the mean CP value of β vector is 0.78 for DGPs with DOF 1 and 12, and these numbers are 0.83 and 0.90 for the error-covariance matrix $\overline{\Sigma}$.[14] Similarly, as expected, the power value of all elements of the γ vector is 1 for both DGPs. The estimator also provides very low variance in power values across the vector β with the mean values of 0.97 and 0.98 for both DGPs. However, the power variance across error-covariance matrix ($\overline{\Sigma}$) slightly increases with the increase in DOF from 1 to 12, but the mean power value decreases from 0.80 to 0.72. Importantly, the power value of all diagonal elements of $\overline{\Sigma}$ is 1 across both DGPs, but it is relatively lower for the off-diagonal elements.

4.5.2 Effect of modeling fat-tailed data with normal distribution

We conduct another Monte Carlo study to understand the consequences of assuming a normally-distributed error kernel when the actual DGP has a fat-tailed error distribution. To accomplish this, we adopt the same model specification as described in the first Monte Carlo study, but consider a DGP with the DOF of 2 to represent a fat-tailed distribution. We generate 50 resamples using this DGP, and estimate the GCM-t model with the fixed DOF of 2 (correct model specification) and the GCM-t model with the fixed DOF of 300 (which is equivalent to

[14] CP of the DOF parameter (δ) in DGP-I is 0.43 (see table 4.2). Even if the estimated DOF is close to the true DOF (low APB), such low CP occurs due to low standard errors.

the GCM-N model) for each resample. To benchmark this analysis, we further generate 50 resamples with a thin-tailed error distribution (DOF value of 12) and estimate GCM-t with DOF values of 12 and 300. The estimation results for DGPs with DOF values of 2 and 12 are provided in Tables 4.4 and 4.5, respectively.

We obtain several insights from this second simulation study. First, modeling thick/heavy-tailed data (DOF <= 7) with a normally-distributed error kernel can introduce a large bias in the parameter estimates. The mean APB value (across all parameters) of the GCM-t model with the fixed DOF of 300 (equivalent to GCM-N model) increases from 14.46% to 88.94% with the decrease in DOF of the DGP from 12 to 2. These values are 11.36% and 11.59% for the GCM-t model with the correct error specification. Second, we also observe a high degree of deterioration in the goodness of fit due to error misspecification. Under the fat-tailed DGP (DOF=2), the loglikelihood value of the GCM-t model with the DOF 2 is 505 points higher than that of the GCM model with DOF of 300 (Table 4.4). This difference in the loglikelihood values at convergence is just around 17 points under the DGP with the DOF of 12.

We also replicate this analysis for the true DGP with DOF=1. When we estimate GCM-t with DOF=300 (i.e., equivalent GCM-N model) on resamples of this DGP, some elements of the error-covariance matrix explode and also result into the wrong direction of the parameter estimates, even when only diagonal elements are estimated. This behavior might be a manifestation of the GCM-N model's attempt to fit the index function and then broaden the support of the

normal distribution around the fitted index function to put the probability mass on tails to mimic the underlying DGP. This phenomenon can potentially explain the common convergence concerns faced by researchers in estimating the error-covariance matrix of probit-based choice models. Thus, problems in recovering the error-covariance matrix can indicate the possibility of a misspecified error kernel (i.e. the underlying assumption of normally distributed error kernel might be incorrect, and a t-distributed error kernel may perform better).

4.6 Empirical study

We now present statistical and behavioral insights from comparison of the GCM-t and GCM-N models in an empirical setting that jointly models individual's vehicle-miles-traveled (VMT) and stated vehicle purchase preferences. We also illustrate how decision-uncertainty behavior of different demographic groups can be captured by the GCM-t model.

4.6.1 Data description

We conducted a stated preference survey of 1542 individuals in 2018 to examine the policies related to on-street parking with charging facilities for battery electric vehicles (EV) in the city of Philadelphia, Pennsylvania. Philadelphia has substantial variation in residents' socioeconomic attributes, driving patterns, and neigh-

borhood characteristics. The city also has the kinds of short trips and stop-and-go traffic that are well-suited to EVs. Charging locations and neighborhood parking, however, remain a substantial barrier to the adoption of EVs. In 2017, the City Council put a moratorium on a ten-year old policy to permit on-street EV charging stations and parking spaces after installing fewer than one hundred of them throughout the city.

The main goal of the survey was to better understand preferences for adoption of electric vehicles by Philadelphians, within the context of availability of dedicated public parking with charging stations. We thus asked survey respondents to imagine themselves in a situation where they had to buy a new vehicle, had settled on a make and model, and must make a choice about whether to buy an electric version, gasoline version, or no car replacement (opt-out option) given a set of vehicle attributes. In a discrete choice experiment (DCE), each survey participant responded to purchase preferences in eight choice situations based on varying purchase prices, operating costs, electric vehicle performance in terms of driving range, and EV parking characteristics (monthly cost for access, time to recharge the battery as a proxy for type of the charging station, average time to find a parking spot as a proxy for availability, and on/off-street location). Table 4.6 presents a sample of a choice situation.

The experimental shares of gasoline vehicles, electric vehicles, and the opt-out alternative in the sample are 64%, 32%, and 4%, respectively. The sample summary statistics of key socio-demographic and built-environment attributes

are reported in Table 4.7. Comparing our survey sample to Census micro data of persons over 17 in households with one or more cars, our respondents are substantially more likely to be female (70% vs 52%), white non-Hispanic (64% vs. 36%), younger (39 vs. 44 years old on average), and well-educated (61% vs. 34% with a BA or higher). The income and housing type of respondents are generally representative of Philadelphia's adult population with cars and those who commute to work by car. For example, 63% live in row homes and 20% live in multi-unit buildings. Although, we had respondents from all Philadelphia Zip Codes, respondents are disproportionately from predominantly white neighborhoods in Northeast Philadelphia. In addition to the vehicle choice experiment, respondents were also asked to report their annual household vehicle-miles-traveled (VMT). In the empirical analysis, we jointly model the household's annual VMT and choice of vehicle as a function of socio-demographic, built-environment, and alternative-specific attributes.

4.6.2 Results and discussion

Table 4.8 and 4.9 summarizes the results of GCM-t and GCM-N models, which we discuss in the next subsections.

Socio-demographic and built-environment factors

In this section, we discuss the relation of individual's or household's characteristics with household's annual VMT and vehicle purchase preferences. Several demographic characteristics have a significant association with the household's VMT in both GCM-t and GCM-N models. The results indicate that married households tend to have lower VMT as compared to unmarried households, ceteris paribus. This association can be a manifestation of the way different types of households spend time on their activities. For example, whereas married (or larger) households are likely to spend more time together at home in joint activities (Fang, 2008; Spissu et al., 2009), unmarried individuals may drive more frequently for various leisure activities (e.g., going to a club, theater, or eating out) besides commuting trips.

As expected, a household with a full-time working individual is likely to drive more than those with part-time or non-workers (Lee and McDonald, 2003; McQuaid and Chen, 2012). Surprisingly, households in higher density Zip Codes tend to drive more than those in lower density ones. This stands in sharp contrast with findings from recent meta-analyses (Ewing and Cervero, 2010; Stevens, 2017), as well as findings from the Philadelphia region more specifically (Klein et al., 2018). The counter-intuitive finding may relate to our sample excluding non-drivers and drawing disproportionately from higher-educated, white, women drivers in relatively dense neighborhoods of Northeast Philadelphia that are somewhat far from Philadelphia's major employment centers.

We now discuss the factors determining the vehicle purchase preferences of Philadelphia residents. We discuss the effect of socio-demographic attributes, followed by built-environment variables. First, married households with children are less inclined toward the purchase of EVs as compared to single individuals, possibly due to the compact size of most current electric cars. Second, households with highly educated individuals (Masters or post-graduate degree) are more likely to purchase electric vehicles as compared to less educated households, perhaps as a sign of highly responsible behavior toward the planet. The literature also confirms that highly educated individuals exhibit a higher degree of environmental consciousness in terms of recycling, purchase of organic food, and awareness about factors contributing to global warming (Fisher et al., 2012). Third, the number of driving license holders has a positive association with the likelihood of purchasing an EV. This is because households with more drivers can better perceive the long-term benefits of buying fuel-efficient EVs as they are likely to have higher VMT. Fourth, as expected, households with a higher number of vehicles are less likely to purchase EVs. This is perhaps because car-lovers generally have higher income and may view compact design and lower range of EVs as major barriers to adoption (Jakobsson et al., 2016). Fifth, households who already own a hybrid EV also exhibit a higher inclination toward purchasing an EV. Sixth, male respondents are more likely to purchase EVs than females. Whereas most studies have found that females are more environmental-friendly than males, some studies have also reported males to be greener than females (Fisher et al., 2012). Seventh, the results indicate that Asians are more likely to purchase EVs followed

by Caucasians. However, low-income groups with generally less access to opportunities (such as African-American and Hispanics) are more likely to stick to traditional gasoline vehicles (Muehlegger and Rapson, 2018). Finally, as expected, millennials are more likely to purchase EVs than baby boomers and generation-X individuals. Millennials grew up in an era of prevalent information on topics such as global warming and environmental sustainability, which perhaps made them more inclined to adopt technology aimed at benefiting the environment.

Both built-environment (walk-score and population density of the neighborhood) factors have expected and intuitive directions of effects on vehicle purchase preferences. In terms of structural effects, households with a higher vehicle mileage are less likely to purchase EVs, possibly due to limitations in driving range or lower concern for the environment.

Willingness to pay estimates

We now analyze the results related to alternative-specific explanatory variables in the choice model. The directions of effects are the same and intuitive in both GCM-t and GCM-N models. Since the magnitude of marginal-utility parameters is not directly comparable, we derive willingness to pay (WTP) for improving various EV attributes.

Figure 4.2 presents the maximum WTP as premium for the purchase of an

electric car with a marginal improvement in driving range.[15] The results of GCM-t and GCM-N models indicate that Philadelphians are willing to pay additional $212 (GCM-t) and $153 (GCM-N) in purchase price, respectively, to increase driving range by one mile for a car offering 150 miles with a full battery. These WTP estimates decrease to $64 (GCM-t) and $46 (GCM-N) when the driving range of the car reaches 500 miles, which is close to parity with gasoline cars. Clearly, GCM-t estimates are higher than those of GCM-N, but the difference decreases with the increase in driving range.

Since monthly cost of having access to public parking with EV charging was an experimental attributes, we also estimate WTP measures for characteristics of the parking spot. For instance, we derive the WTP to reduce average search time for available EV parking. According to GCM-t and GCM-N estimates, Philadelphians are willing to add $6.7 and $4.6 in their monthly parking cost to reduce parking search time by a minute, respectively. This conforms to findings that on-street residential parking is frequently underpriced (Shoup, 2005). In a neighborhood where drivers spend just five minutes searching for parking on average, the average stated willingness to pay to avoid that search time is eight to eleven times higher than the price of a parking permit for a single vehicle.

WTP to reduce the EV charging time by an hour ranges from $96 (GCM-t) to $60 (GCM-N). The higher WTP estimates of GCM-t are aligned with the finding of the study by Dekker et al. (2016), who also observed an increase in WTP for

[15] Since driving range enters the utility specification logarithmically, the WTP estimates are non-linear and decreases with the driving range.

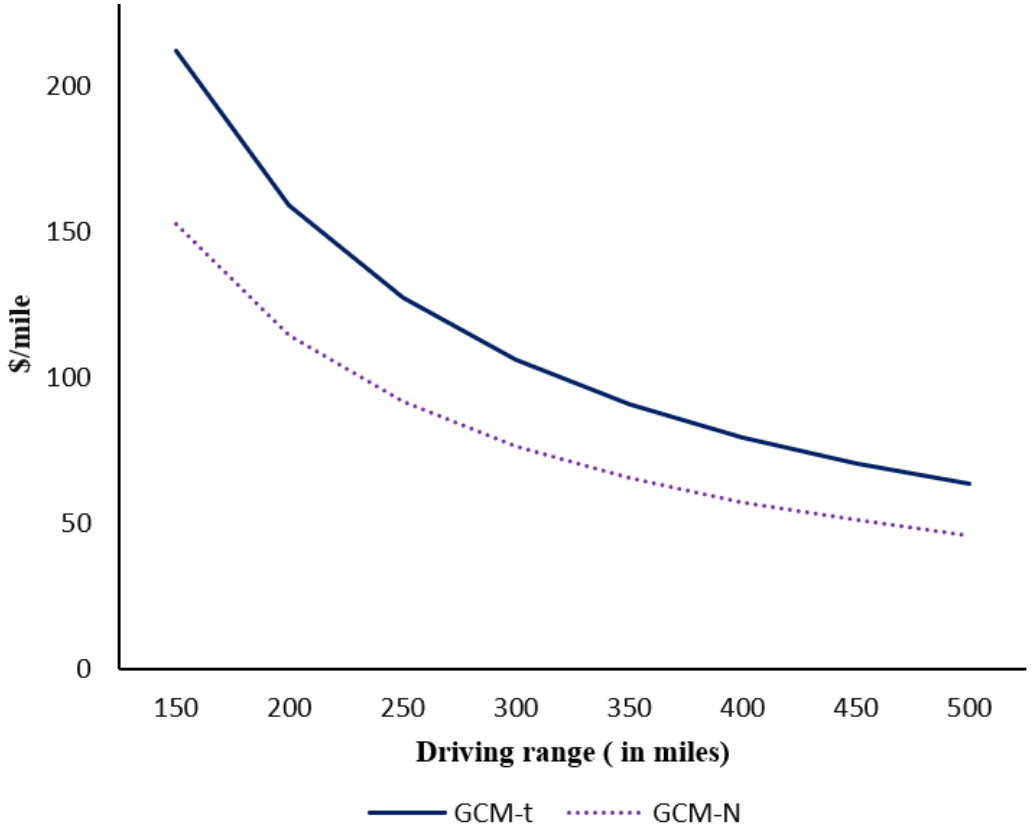

Figure 4.2: Willingness to pay to increase the driving range of an electric vehicle by a mile

flood risk reductions after controlling for the behavioral responses to decision uncertainty in an integrated choice and latent variable model.

Finally, because vehicle purchase and parking costs happen at different times, the annual subjective discount rate of parking cost is estimated at 10.7% (GCM-t) and 9.3% (GCM-N), which is slightly above market interest rates in the automotive industry.

Behavioral Insights

We parameterize the DOF of the t-distributed error kernel as a function of demographics in GCM-t to capture decision-uncertainty behavior of decision-makers. Table 4.10 presents the relation between the DOF and demographics in GCM-t, which is used to obtain the DOF of each respondent. These results also provide insights about decision uncertainty of different socio-demographic groups. For example, the positive relationship between married males and the DOF indicates that married males are likely to have a higher DOF, and thus are less certain about their choices as compared to unmarried females.

We also compute elasticity estimates with respect to 1% and 25% reduction in the EV parking price for both models. Tables 4.11 and 4.12 present elasticity estimates for 1% and 25% reduction, respectively, for ten different ranges of the DOF such that each bin contains 10% respondents of the sample.[16] Whereas the difference between elasticity estimates of GCM-t and GCM-N models are not apparent for the 1% reduction, they are quite stark for the 25% reduction in parking price (see Table 4.12). In general, the magnitude of elasticity estimates for gasoline and electric alternatives in the GCM-N model is higher (1.5 times, on average) than that of the corresponding GCM-t model. Specifically, this ratio is higher for respondents with lower DOF values. This supports the our earlier observation that individuals with lower DOF are more certain about their choices and are less

[16]Note that there is no concept of DOF in the GCM-N model. To facilitate the comparison between GCM-t and GCM-N models, DOF ranges are obtained from the GCM-t model and then the same set of individuals are used for respective calculations in the GCM-N models.

sensitive to changes in the utility of alternatives.

Finally, we plot the change in the probability of choosing electric and gasoline vehicles in GCM-t and GCM-N model as a function of the change in the utility of electric vehicle in Figures 4.3 and 4.4, respectively. For illustration, we only make plots for three DOF ranges. These plots offer many interesting insights. First, plots for three DOF ranges in the GCM-N model are virtually the same. Second, these change-in-probability plots become steeper with the increase in the DOF in the GCM-t model, capturing the varied decision-uncertainty behavior in the sample. Finally, curves for the lower DOFs are not asymptotic to the x-axis, demonstrating a phenomenon that it is implausible to achieve 100% market share for any alternative in spite of being far superior to other available options.

Prediction performance

The dataset used in this case study is a good example of class-imbalance as the opt-out alternative has a very small sample share. This is an appropriate scenario to compare the predicted probability of the chosen alternative in GCM-t and GCM-N model. Table 4.12 presents the ratio of GCM-t and GCM-N choice probabilities for a group of respondents with a varying range of DOF.

The probability ratios for two alternatives (Gasoline and Electric vehicle which have 96% sample share in total) are close to 1 across all DOF ranges. However, the average ratio of 1.15 for the opt-out option (4% sample share) confirms our earlier

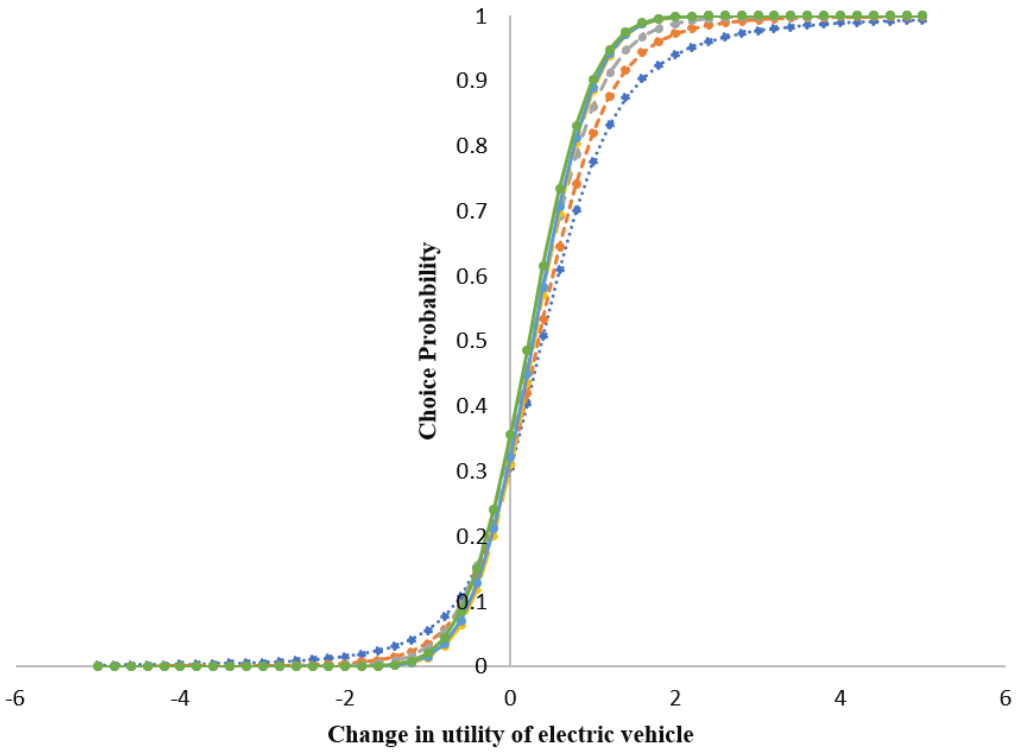

Figure 4.3: Probability of choosing electric vehicle due to change in utility of electric vehicle.

observation that GCM-t model is indeed better than a GCM-N model in modeling such imbalanced datasets. These ratios for the opt-out option are significantly higher than 1 in lower ranges of the DOF (≤ 3.79). This finding is aligned with our discussion in subsection 4.4.1 on the ability of GCM-t to better predict the preference for the low-share alternatives under fat-tailed error distributions.

Surprisingly, GCM-N outperforms GCM-t for DOF ranges 4.16-4.40 and 5.33-9.32 with probability ratios 0.77 and 0.80, respectively. Since GCM-t can approximate GCM-N with the higher DOF, we would not expect this pattern. We speculate that such anomalies might be a manifestation of the strict constraint to use

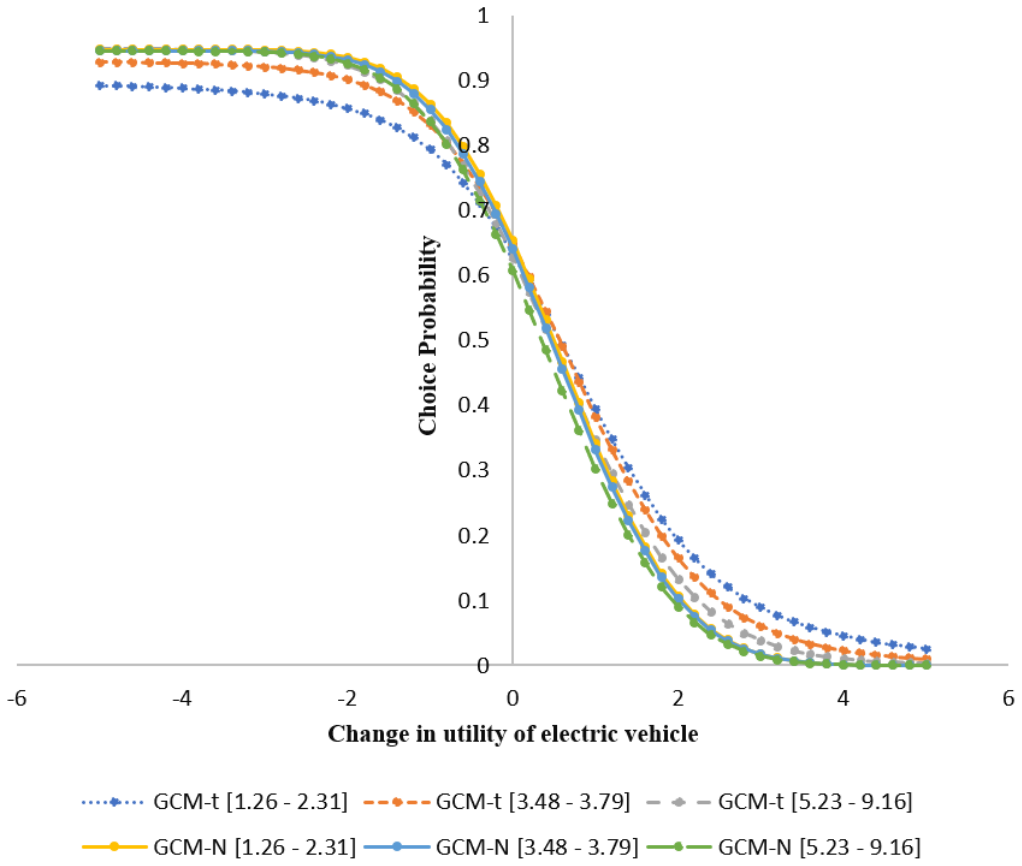

Figure 4.4: Probability of choosing gasoline vehicle due to change in utility of electric vehicle.

the same DOF expression across all alternatives of the choice model and the continuous response model. Nonetheless, the superior performance of GCM-N over GCM-t for certain DOF values can be leveraged by developing a latent class model with two classes where classes differ based on the distributional assumptions on the error kernel (i.e., t and normal). We discuss another flexible way to relax the "same DOF" assumption in section 4.7.

Model selection

We first compare the goodness-of-fit measures of GCM-t and GCM-N models. The log-likelihood value at convergence of GCM-t is better than the corresponding GCM-N model by 200 points (see Table 4.9). Further, the lower Bayesian information criterion (BIC) value of GCM-t also suggests that the DGP of this empirical study is better modeled using the GCM-t model (with 39 parameters) than the GCM-N model (with 42 parameters).

Apart from these measures, we also compare the trace of the error-covariance matrix of GCM-t and GCM-N to assess the amount of (un)explained variance in these models (see Table 4.11). The trace values of the error-covariance matrix in GCM-t and GCM-N models are 1.06 and 3.44, respectively. This result is as expected because parameterization of the DOF in GCM-t model enables it to explain more variation in the DGP than the GCM-N model.

Piatek and Gensowski (2017) did a similar effort to reduce the unexplained variance in the MNP model with latent factors. The authors split the error-covariance matrix into two parts – the allocation (or factor loading) matrix which can be specified to model a covariance structure across alternatives, and the idiosyncratic error matrix which captures the correlation in residual. The key idea behind the allocation matrix is to allow the researcher to specify driving factors behind the choice variation beyond the observables included in the deterministic part of the utility. However, the allocation matrix is not straightforward to

define in empirical studies and, beyond a mathematical inconvenience, is difficult to parametrize as a function of other variables. Thus, in contrast to splitting the error-covariance matrix, parametrization of the DOF in the GCM-t model not only offers insights about choice behavior but also provides a new parsimonious specification to reduce the unexplained variance.

4.7 Conclusions and future work

In this study, we have proposed and applied for the first time a t-distributed error kernel within random-utility-maximization (RUM) multinomial choice models. We have not only discussed statistical advantages of using this t-distributed kernel over a normally-distributed one in class-imbalance situations, but also illustrated how the t-distributed error kernel can capture decision-uncertainty behavior (i.e., how certain the decision-makers are about the choices that they make). Furthermore, we have extended this model to a generalized continuous-multinomial response model with a t-distributed error kernel (GCM-t). Using the composite marginal likelihood method, separation-of-variables approach, and properties of the t-distribution, we have derived the full-information maximum likelihood estimator of the GCM-t model and tested its statistical properties in a Monte Carlo study. Finally, we have compared the statistical performance of GCM-t and GCM with normally-distributed error kernel (GCM-N) in a simulation study, and validated behavioral hypotheses about the advantages of GCM-t

over GCM-N in an empirical study related to the adoption of electric vehicles in the city of Philadelphia.

The results of the simulation study indicate that GCM-t and GCM-N perform equally well in terms of recovering model parameters, goodness-of-fit, and model inference when the true error distribution is thin-tailed. However, GCM-t either outperforms GCM-N by a significant margin or GCM-N may not even converge in datasets with fat-tailed error distributions. In the case study about adoption of electric vehicles, GCM-t is more accurate than GCM-N in estimating the choice of the alternative with a small sample share. Moreover, accounting for decision-uncertainty behavior in GCM-t manifests into lower elasticity estimates (relative to those of GCM-N). These lower elasticity estimates of GCM-t, as well as the evidence we found for underestimation in the GCM-N model of individual's willingness to pay to increase the driving range of electric vehicles and for reducing the parking search time for a parking spot with charging are all relevant in the context of planning policies for broader adoption of electric vehicles.

We now discuss key limitations of this study and possible research avenues to overcome them in the future *First*, unlike GCM-N, incorporating random parameters with parametric heterogeneity distributions in GCM-t is possible, but not straightforward because the resulting distribution – the sum of t-distributed random variables or sum of t-distributed random variable and other parametric distributions – is not of a known form. Therefore, even inclusion of parametric unobserved preference heterogeneity in the GCM-t model requires an additional layer

of simulation to compute the likelihood function of the model. This necessity of adding a simulation layer would motivate researchers to rather incorporate flexible semi-parametric (instead of parametric) preference heterogeneity distributions in the GCM-t model (see Train, 2016; Vij and Krueger, 2017; Bansal et al., 2018*c*, for recent developments in semi-parametrics). We leave this extension for future work. *Second*, we considered the same degrees of freedom across all alternatives of the choice model and the continuous variable model because the exact distribution of linear or non-linear combinations of two t-distributed random variables with a arbitrary degrees-of-freedom values is not known. Extending GCM-t to a more flexible model with varying degrees of freedom across alternatives using a copula-based approach (Bhat and Eluru, 2009) is a potential direction for future developments. This copula-based extension would allow researchers to capture decision-uncertainty behavior across alternatives, not just across decision-makers. *Third*, the proposed GCM-t model cannot handle asymmetric error distributions. From an estimation standpoint, extending GCM-t model to a GCM model with a skew-t-distributed error kernel is not challenging because all relevant statistical properties are well-established in the literature (Azzalini and Genton, 2008). However, practical concerns related to model inference and asymptotic properties of this new estimator require further investigation (Azzalini and Arellano-Valle, 2013).

Finally, the full-information maximum likelihood estimation of GCM-t is slow because it relies on a numerical gradient in the absence of tractable expressions of the analytical gradient and Hessian matrix. The possibilities of developing a

Bayesian estimator (Kim et al., 2007) or an expectation-maximization procedure (Liu, 2004) can be explored in the future to speed up the estimation of the GCM-t model.

Table 4.2: Simulation Results of GCM-t Model for DOF-I scenario (DOF=1)

Parameter	True Value	MEV	MAB	APB	FSSE	ASE	ASE/ FSSE	CP	Power
γ_{11}	1	0.98	0.02	1.69	0.06	0.06	1.07	0.93	1
γ_{12}	0.5	0.51	0.01	2.06	0.06	0.06	1.12	0.97	1
γ_{13}	0.75	0.76	0.01	1.46	0.07	0.06	0.98	0.93	1
γ_{14}	-0.5	-0.5	0	0.94	0.06	0.06	1.06	0.97	1
β_{21}	-1.5	-1.74	0.24	16.23	0.28	0.16	0.56	0.47	1
β_{22}	1	0.91	0.09	8.71	0.11	0.11	0.99	0.89	1
β_{23}	0.9	0.81	0.09	10.35	0.13	0.12	0.98	0.9	1
β_{25}	1	0.94	0.06	6.46	0.26	0.16	0.63	0.72	0.99
β_{31}	-1.3	-1.47	0.17	13.21	0.26	0.2	0.77	0.85	1
β_{32}	0.9	0.79	0.11	12.35	0.14	0.12	0.87	0.81	1
β_{33}	0.8	0.68	0.12	15.58	0.16	0.13	0.84	0.79	1
β_{35}	0.9	0.8	0.1	10.9	0.23	0.17	0.73	0.8	0.99
β_{41}	-1.2	-1.36	0.16	13.49	0.27	0.2	0.76	0.79	1
β_{42}	0.8	0.67	0.13	16.54	0.15	0.13	0.88	0.76	1
β_{43}	0.7	0.57	0.13	18.7	0.19	0.15	0.81	0.8	0.93
β_{45}	0.8	0.68	0.12	15.23	0.25	0.18	0.74	0.77	0.92
β_{51}	-1	-1.08	0.08	7.72	0.22	0.18	0.83	0.91	1
β_{52}	0.7	0.54	0.16	23.33	0.15	0.12	0.85	0.65	0.99
β_{53}	0.6	0.43	0.17	27.66	0.17	0.15	0.84	0.76	0.83
β_{55}	0.7	0.56	0.14	20.14	0.23	0.18	0.81	0.81	0.8
Covariance matrix elements									
$\overline{\Sigma}_{11}$	1.5	1.62	0.12	8.28	0.1	0.09	0.97	0.75	1
$\overline{\Sigma}_{21}$	0.3	0.28	0.02	8.22	0.29	0.18	0.61	0.77	0.42
$\overline{\Sigma}_{31}$	0.4	0.44	0.04	9.78	0.3	0.2	0.66	0.79	0.6
$\overline{\Sigma}_{33}$	1.1	1.21	0.11	10.41	0.29	0.24	0.83	0.89	1
$\overline{\Sigma}_{41}$	0.6	0.72	0.12	20.24	0.35	0.24	0.7	0.78	0.76
$\overline{\Sigma}_{44}$	1.2	1.48	0.28	23.19	0.4	0.33	0.83	0.91	1
$\overline{\Sigma}_{51}$	0.5	0.6	0.1	19.12	0.34	0.24	0.72	0.82	0.65
$\overline{\Sigma}_{55}$	1.3	1.48	0.18	14.06	0.41	0.31	0.76	0.9	1
Degree-of-freedom									
δ	1	1.08	0.08	8.24	0.03	0.04	1.1	0.43	1
Average			**0.14**	**12.56**	**0.21**	**0.16**	**0.84**	**0.8**	**0.93**

Note1: MEV: mean estimated value, MAB: mean absolute bias, APB: absolute percentage bias.
Note2: FSSE: finite sample standard error, ASE: asymptotic standard error, CP: coverage probability.

Table 4.3: Simulation Results of GCM-t Model for DOF-II scenario (DOF=12)

Parameter	True Value	MEV	MAB	APB	FSSE	ASE	ASE/ FSSE	CP	Power
γ_{11}	1	1	0	0.38	0.04	0.05	1.09	0.97	1
γ_{12}	0.5	0.5	0	0.77	0.05	0.05	0.99	0.97	1
γ_{13}	0.75	0.75	0	0.42	0.05	0.05	0.94	0.91	1
γ_{14}	-0.5	-0.5	0	0.26	0.05	0.05	0.92	0.91	1
β_{21}	-1.5	-1.76	0.26	17.62	0.17	0.15	0.88	0.49	1
β_{22}	1	0.89	0.11	11.48	0.11	0.11	0.96	0.8	1
β_{23}	0.9	0.79	0.11	12.12	0.12	0.12	0.98	0.87	1
β_{25}	1	0.9	0.1	10.28	0.18	0.19	1.08	0.94	1
β_{31}	-1.3	-1.42	0.12	9.58	0.27	0.27	1.01	0.95	1
β_{32}	0.9	0.76	0.14	16.03	0.13	0.13	1.02	0.77	1
β_{33}	0.8	0.67	0.13	16.38	0.15	0.14	0.94	0.77	1
β_{35}	0.9	0.74	0.16	17.89	0.18	0.19	1.06	0.83	0.99
β_{41}	-1.2	-1.29	0.09	7.49	0.25	0.26	1.04	0.97	1
β_{42}	0.8	0.63	0.17	21.82	0.13	0.12	0.92	0.65	1
β_{43}	0.7	0.53	0.17	24.39	0.16	0.13	0.83	0.62	0.98
β_{45}	0.8	0.64	0.16	20.04	0.19	0.19	1.04	0.87	0.93
β_{51}	-1	-1.02	0.02	1.86	0.2	0.22	1.11	0.98	1
β_{52}	0.7	0.5	0.2	28.48	0.12	0.11	0.94	0.52	1
β_{53}	0.6	0.42	0.18	30.18	0.14	0.12	0.88	0.61	0.92
β_{55}	0.7	0.53	0.17	24.78	0.18	0.19	1.02	0.79	0.81
Covariance matrix elements									
$\overline{\Sigma}_{11}$	1.5	1.52	0.02	1.22	0.06	0.06	1.05	0.97	1
$\overline{\Sigma}_{21}$	0.3	0.17	0.13	43.74	0.21	0.19	0.9	0.85	0.19
$\overline{\Sigma}_{31}$	0.4	0.34	0.06	15.32	0.23	0.2	0.87	0.9	0.41
$\overline{\Sigma}_{33}$	1.1	1.1	0	0.04	0.41	0.41	1.01	0.93	1
$\overline{\Sigma}_{41}$	0.6	0.55	0.05	8.92	0.29	0.23	0.81	0.87	0.67
$\overline{\Sigma}_{44}$	1.2	1.21	0.01	0.53	0.48	0.48	1.01	0.9	1
$\overline{\Sigma}_{51}$	0.5	0.43	0.07	13.6	0.26	0.22	0.84	0.89	0.49
$\overline{\Sigma}_{55}$	1.3	1.2	0.1	7.58	0.4	0.45	1.1	0.91	1
Degree-of-freedom									
δ	12	13.48	1.48	12.37	2.89	2.84	0.98	0.97	1
Average			**0.18**	**12.95**	**0.28**	**0.27**	**0.97**	**0.84**	**0.91**

Note1: MEV: mean estimated value, MAB: mean absolute bias, APB: absolute percentage bias.
Note2: FSSE: finite sample standard error, ASE: asymptotic standard error, CP: coverage probability.

Table 4.4: Effect of ignoring non-normality (DOF=2)

Parameter	True Value	MEV	MAB	APB	MEV	MAB	APB
		GCM-t with DOF of 2			GCM-t with DOF of 300		
Mean effect elements							
γ_{11}	1	0.98	0.02	1.53	0.98	0.02	1.6
γ_{12}	0.5	0.51	0.01	2.12	0.51	0.01	2.51
γ_{13}	0.75	0.76	0.01	0.99	0.76	0.01	1.93
γ_{14}	-0.5	-0.5	0	0.67	-0.51	0.01	2.48
β_{21}	-1.5	-1.75	0.25	16.91	-1.4	0.1	6.36
β_{22}	1	0.9	0.1	10.13	0.66	0.34	33.79
β_{23}	0.9	0.81	0.09	9.94	0.6	0.3	32.91
β_{25}	1	0.91	0.09	8.95	0.65	0.35	35.18
β_{31}	-1.3	-1.49	0.19	14.34	-1.33	0.03	2.44
β_{32}	0.9	0.76	0.14	16.1	0.55	0.35	38.33
β_{33}	0.8	0.67	0.13	16.66	0.49	0.31	38.6
β_{35}	0.9	0.82	0.08	9.33	0.6	0.3	33.77
β_{41}	-1.2	-1.36	0.16	13.28	-1.16	0.04	2.98
β_{42}	0.8	0.64	0.16	20.45	0.44	0.36	45.38
β_{43}	0.7	0.5	0.2	28.62	0.33	0.37	52.55
β_{45}	0.8	0.71	0.09	10.75	0.5	0.3	37.95
β_{51}	-1	-1.09	0.09	8.82	-2.65	1.65	164.78
β_{52}	0.7	0.53	0.17	24.31	0.34	0.36	51.03
β_{53}	0.6	0.42	0.18	30.29	0.25	0.35	58.77
β_{55}	0.7	0.55	0.15	21.58	0.28	0.42	60.32
Covariance matrix elements							
$\overline{\Sigma}_{11}$	1.5	1.57	0.07	4.42	4.2	2.7	180.26
$\overline{\Sigma}_{21}$	0.3	0.24	0.06	21.37	-0.37	0.67	222.38
$\overline{\Sigma}_{31}$	0.4	0.32	0.08	19.77	-0.18	0.58	145.74
$\overline{\Sigma}_{33}$	1.1	1.09	0.01	0.72	1.54	0.44	39.58
$\overline{\Sigma}_{41}$	0.6	0.55	0.05	9.13	0.31	0.29	48.77
$\overline{\Sigma}_{44}$	1.2	1.23	0.03	2.25	1.39	0.19	15.44
$\overline{\Sigma}_{51}$	0.5	0.51	0.01	1.49	1.07	0.57	113.11
$\overline{\Sigma}_{55}$	1.3	1.37	0.07	5.48	14.58	13.28	1021.4
Degree-of-freedom							
δ	2	2.12	0.12	5.84			
Average			**0.1**	**11.59**		**0.88**	**88.94**
Loglikelihood		-10417.44			-10922.73		

Note: MEV: mean estimated value, MAB: mean absolute bias, APB: absolute percentage bias.

Table 4.5: Effect of ignoring non-normality (DOF=12)

Parameter	True Value	MEV	MAB	APB	MEV	MAB	APB
		GCM-t with DOF of 12			GCM-t with DOF of 300		
Mean effect elements							
γ_{11}	1	1	0	0.39	1	0	0.32
γ_{12}	0.5	0.49	0.01	1.46	0.49	0.01	1.77
γ_{13}	0.75	0.75	0	0.45	0.75	0	0.03
γ_{14}	-0.5	-0.5	0	0.26	-0.5	0	0.01
β_{21}	-1.5	-1.75	0.25	16.83	-1.7	0.2	13.54
β_{22}	1	0.89	0.11	11	0.85	0.15	14.99
β_{23}	0.9	0.81	0.09	10.02	0.78	0.12	13.87
β_{25}	1	0.89	0.11	10.84	0.85	0.15	14.84
β_{31}	-1.3	-1.4	0.1	7.98	-1.36	0.06	4.97
β_{32}	0.9	0.77	0.13	13.98	0.74	0.16	18.16
β_{33}	0.8	0.72	0.08	10.01	0.69	0.11	13.78
β_{35}	0.9	0.71	0.19	20.82	0.68	0.22	24.95
β_{41}	-1.2	-1.32	0.12	10.17	-1.31	0.11	9.15
β_{42}	0.8	0.63	0.17	21	0.6	0.2	25.02
β_{43}	0.7	0.54	0.16	22.19	0.52	0.18	25.77
β_{45}	0.8	0.64	0.16	19.57	0.61	0.19	24.17
β_{51}	-1	-1.03	0.03	2.95	-1.04	0.04	4.31
β_{52}	0.7	0.52	0.18	25.83	0.49	0.21	29.86
β_{53}	0.6	0.45	0.15	25.09	0.42	0.18	29.24
β_{55}	0.7	0.52	0.18	26.39	0.49	0.21	30.27
Covariance matrix elements							
$\overline{\Sigma}_{11}$	1.5	1.51	0.01	0.33	1.77	0.27	17.8
$\overline{\Sigma}_{21}$	0.3	0.2	0.1	32.59	0.15	0.15	50.2
$\overline{\Sigma}_{31}$	0.4	0.4	0	0.16	0.38	0.02	4.28
$\overline{\Sigma}_{33}$	1.1	1.13	0.03	3.15	1.14	0.04	3.63
$\overline{\Sigma}_{41}$	0.6	0.57	0.03	5.12	0.59	0.01	1.82
$\overline{\Sigma}_{44}$	1.2	1.32	0.12	10	1.41	0.21	17.19
$\overline{\Sigma}_{51}$	0.5	0.47	0.03	5.93	0.48	0.02	4.8
$\overline{\Sigma}_{55}$	1.3	1.28	0.02	1.77	1.38	0.08	6.22
Degree-of-freedom							
δ	12	13.56	1.56	13.04			
Average			**0.14**	**11.36**		**0.12**	**14.46**
Loglikelihood		-8977.45			-8994.84		

Note: MEV: mean estimated value, MAB: mean absolute bias, APB: absolute percentage bias.

Table 4.6: Sample of a choice situation in the discrete choice experiment

	Gasoline version	**Electric version**
Purchase Price	$19,000	$34,000
Driving cost per 50 miles	$5.5 per 50 miles	$2.5 per 50 miles
Electric driving range		250 miles
EV parking: charging time for 50 miles		90 minutes per 50 miles
EV parking: charging type		On-street
EV parking: time to find space		5 minutes
EC parking: monthly price		$50 per month

Given the 2 options above. which car would you buy?

- Gasoline version
- Electric version
- Neither

Table 4.7: Descriptive statistics of the sample

Variables	**Frequency/Average**
Married indicator	36.45%
Indicator for having children	49.55%
Indicator for working full time	59.47%
Indicator for holding master's or above degree	17.70%
Number of adults in household	2.88
Number of driving-license-holders in household	2.03
Number of vehicles in household	1.70
Indicator for owning hybrid electric vehicle	4.17%
Male indicator	29.51%
Hispanic indicator	9.53%
Walk score	77.33
Population density	21.53
Household annual vehicle miles traveled	14883.60
Race	
African-American indicator	24.38%
Asian indicator	3.76%
Caucasian indicator	62.06%
Age category	
Baby-boomer indicator	16.41%
GenX indicator	24.58%
Millenial indicator	57.65%

Table 4.8: Comparison of GCM-t and GCM-N in empirical study (Part 1, t-value in parenthesis)

Model fit Statistics	**GCM-t Model**				**GCM-N Model**			
Parameters	**Continuous**	**Unordered**			**Continuous**	**Unordered**		
	Household VMT	**Gasoline**	**Electric**	**Opt-out**	**Household VMT**	**Gasoline**	**Electric**	**Opt-out**
Intercept	1.983		-1.755	-4.229	1.856		-1.599	-5.027
	(81.22)		(-5.69)	(-9.82)	(34.64)		(-5.66)	(-7.83)
Demographic Variables								
Married indicator	-0.059				-0.105		-0.033	
	(-4.38)				(-6.19)		(-1.10)	
Indicator for having children			-0.244		-0.030		-0.227	
			(-7.80)		(-4.50)		(-7.09)	
Indicator for working full time	0.174				0.136			
	(13.38)				(9.16)			
Indicator for holding master's or above degree			0.144		0.025		0.119	
			(3.94)		(1.14)		(3.44)	
Number of adults in household					0.109		0.049	
					(15.05)		(4.15)	
Number of driving license holders in household	0.055		0.076		-0.033			
	(6.86)		(3.75)		(-2.91)			
Number of vehicles in household			-0.076		0.013		-0.068	
			(-3.78)		(1.29)		(-4.14)	
Indicator for owning hybrid electric vehicle			0.580				0.522	
			(8.12)				(8.02)	
Male indicator							0.081	
							(2.85)	
Hispanic indicator			-0.139		-0.061		-0.119	
			(-2.80)		(-2.62)		(-2.77)	
Race (base: Caucasian)								
African-American indicator					0.159		-0.071	
					(9.62)		(-2.25)	
Asian indicator					-0.102		0.274	
					(-2.80)		(4.31)	
Age category (base: Millennial)								
Baby boomer indicator			-0.473				-0.436	
			(-10.30)				(-10.10)	
GenX indicator			-0.207				-0.189	
			(-5.51)				(-5.60)	
Environmental variables								
Walk-score			0.635		0.103		0.538	
			(6.46)		(1.36)		(6.17)	
Population density of neighborhood	0.004				0.003			
	(5.60)				(2.93)			

Note: VMT is vehicle-miles-traveled.

Table 4.9: Comparison of GCM-t and GCM-N in empirical study (Part 2, t-value in parenthesis)

	GCM-t Model				GCM-N Model			
	Continuous	Unordered			Continuous	Unordered		
	Household VMT	Gasoline	Electric	Opt-out	Household VMT	Gasoline	Electric	Opt-out
Alternative specific variables								
Price (in $1,000)		-0.037	-0.019			-0.037	-0.021	
		(-6.41)	(-3.48)			(-6.77)	(-4.13)	
Operating cost per 50 miles (in $)		-0.284	-0.426			-0.245	-0.328	
		(-7.16)	(-6.97)			(-6.57)	(-6.18)	
Log of driving range (in 100 miles)			0.604				0.481	
			(8.29)				(7.50)	
Electric vehicle charging time (in hours)			-0.215				-0.169	
			(-7.37)				(-6.37)	
Electric vehicle parking search time (in minutes)			-0.015				-0.013	
			(-3.00)				(-2.99)	
Monthly electric vehicle parking cost (in $100)			-0.224				-0.282	
			(-3.05)				(-4.39)	
Structural effect								
Household vehicle miles traveled (in 1000 miles)			-0.019				-0.007	
			(-6.77)				(-4.85)	
Model fit statistics								
Sample size	12336				12336			
Number of parameters	39				42			
Loglikelihood	-22725.81				-22929.19			
Bayesian Information Criterion	45611.18				46030.21			
Trace of covariance matrix	1.06				3.43			

Note: VMT is vehicle-miles-traveled.

Table 4.10: Degree-of-freedom specification results in GCM-t model (t-value in parenthesis)

Parameter	**Estimates (T-value)**
Intercept	1.191 (4.67)
Male indicator	0.202 (2.25)
Married indicator	0.128 (1.46)
Indicator for holding master's or above degree	0.182 (1.67)
Number of adults in household	-0.165 (-5.32)
Number of driving license holders in household	0.132 (2.65)
Walk score	0.372 (1.30)
Race (base: Caucasian)	
African-American indicator	-0.396 (-4.64)
Asian indicator	-0.309 (-1.48)

Table 4.11: Change in choice probability due to 1% reduction in parking-cost of electric vehicle

DOF lower limit	**DOF upper limit**	**GCM-t model**			**GCM-N model**		
		Gasoline	**Electric Vehicle**	**opt-out**	**Gasoline**	**Electric Vehicle**	**opt-out**
1.27	2.37	-0.0002	0.0001	0.0000	-0.0003	0.0003	0.0000
2.37	2.83	-0.0002	0.0002	0.0000	-0.0003	0.0002	0.0000
2.83	3.16	-0.0002	0.0002	0.0000	-0.0003	0.0003	0.0000
3.16	3.53	-0.0002	0.0002	0.0000	-0.0003	0.0002	0.0000
3.53	3.86	-0.0002	0.0002	0.0000	-0.0003	0.0003	0.0000
3.86	4.16	-0.0002	0.0002	0.0000	-0.0003	0.0003	0.0000
4.16	4.40	-0.0002	0.0002	0.0000	-0.0004	0.0003	0.0000
4.40	4.83	-0.0002	0.0002	0.0000	-0.0004	0.0003	0.0000
4.83	5.33	-0.0003	0.0002	0.0000	-0.0004	0.0003	0.0000
5.33	9.32	-0.0003	0.0002	0.0000	-0.0004	0.0003	0.0000

Table 4.12: Change in choice probability due to 25% reduction in parking-cost of electric vehicle

DOF lower limit	DOF upper limit	GCM-t model			GCM-N model		
		Gasoline	Electric Vehicle	opt-out	Gasoline	Electric Vehicle	opt-out
1.27	2.37	-0.0050	0.0038	-0.0005	-0.0087	0.0064	-0.0004
2.37	2.83	-0.0054	0.0041	-0.0006	-0.0088	0.0061	-0.0006
2.83	3.16	-0.0052	0.0044	-0.0005	-0.0084	0.0068	-0.0005
3.16	3.53	-0.0053	0.0043	-0.0004	-0.0084	0.0062	-0.0004
3.53	3.86	-0.0056	0.0050	-0.0006	-0.0089	0.0073	-0.0005
3.86	4.16	-0.0057	0.0051	-0.0005	-0.0088	0.0070	-0.0006
4.16	4.40	-0.0062	0.0053	-0.0004	-0.0094	0.0072	-0.0006
4.40	4.83	-0.0060	0.0052	-0.0007	-0.0091	0.0070	-0.0009
4.83	5.33	-0.0072	0.0053	-0.0004	-0.0104	0.0073	-0.0005
5.33	9.32	-0.0069	0.0046	-0.0005	-0.0100	0.0064	-0.0007

Table 4.13: Ratio of GCM-t and GCM-N probabilities for chosen alternative for different DOF

DOF lower limit	DOF upper limit	Gasoline	Electric Vehicle	opt-out	Sample size
1.27	2.37	0.96	0.96	1.91	1232
2.37	2.83	0.97	1.00	1.33	1232
2.83	3.16	0.99	0.96	1.30	1232
3.16	3.53	0.98	1.00	1.18	1208
3.53	3.86	1.01	0.97	1.44	1264
3.86	4.16	1.00	1.00	0.98	1216
4.16	4.40	1.01	1.00	0.77	1248
4.40	4.83	1.02	1.00	0.87	1224
4.83	5.33	1.02	1.01	0.96	1240
5.33	9.32	1.03	1.00	0.80	1232
Overall Average		**1.00**	**0.99**	**1.15**	

Table 4.14: Covariance matrix (t-value in parenthesis)

GCM-t model		
0.330		
(40.42)		
0.109	1.0	
(7.12)	(fixed)	
-0.020	0.250	0.728
(-1.31)	(2.35)	(3.16)

GCM-N model		
0.638		
(84.77)		
0.109	1.0	
(5.42)	(fixed)	
0.029	0.348	2.790
(1.07)	(1.58)	(2.60)

APPENDIX A

APPENDIX OF CHAPTER 1

A.1 Model Specification: Willingness to Pay Space

Consider a standard discrete choice setting where individual $n \in \{1, \ldots, N\}$ chooses one alternative from the mutually exclusive choice set $\{1, \ldots, J\}$ (indexed by j) over the set of discrete time periods $\{1, \ldots, T\}$ or choice situations (indexed by t). The random utility maximization model is specified as

$$U_{njt} = \mathbf{x}_{njt}'\boldsymbol{\zeta}_n + \varepsilon_{njt} = \gamma_n^R \left(\rho_{njt} + \left[\mathbf{x}_{njt}^{F\prime}, \mathbf{x}_{njt}^{R\prime} \right] \begin{bmatrix} \boldsymbol{\delta}^F \\ \boldsymbol{\delta}_n^R \end{bmatrix} \right) + \varepsilon_{njt} \tag{A.1}$$

where[1] U_{njt} is the random indirect utility associated with individual n choosing alternative j during choice situation t, and ε_{njt} is an iid extreme value type I preference shock. Moreover, both the alternative attributes and WTP parameters are sorted in two groups. On the one hand, $\boldsymbol{\delta}^F$ is a vector of fixed WTP and $\mathbf{x}_{njt}^F$ is the attribute/covariate vector associated with these fixed WTP. On the other hand, $\boldsymbol{\delta}_n^R$ is a vector of random parameters and $\mathbf{x}_{njt}^R$ is the attribute vector for which the researcher expects the presence of unobserved preference heterogeneity. γ_n^R is marginal utility of price ρ_{njt}, which is assumed to be random. The mixing distribution of the set of random parameters $\boldsymbol{\eta}_n^R = \{\gamma_n^R, \boldsymbol{\delta}_n^R\}$ is modeled semi-parametrically below.

If i_{nt} denotes the alternative observed to be chosen by individual n at time

[1] We use negative of price in equation A.1 to get correct sign on WTP estimates.

t, consider now the sequence of chosen alternatives for the decision maker $\{i_{n1}, \ldots, i_{nT}\}$. The probability that individual n made this sequence of choices, conditional on $\boldsymbol{\zeta}_n$, is:

$$L_n(\boldsymbol{\zeta}_n) = \prod_{t=1}^{T} Q_{ni_{nt}t}(\boldsymbol{\zeta}_n) \tag{A.2}$$

where $Q_{ni_{nt}t}(\boldsymbol{\zeta}_n)$ is the probability of individual n choosing alternative i_{nt} in choice situation t. The conditional choice probability $Q_{ni_{nt}t}(\boldsymbol{\zeta}_n)$ is given by the following conditional logit expression:

$$Q_{ni_{nt}t}(\boldsymbol{\zeta}_n) = \frac{e^{U_{ni_{nt}t}}}{\sum_{j=1}^{J} e^{U_{njt}}}. \tag{A.3}$$

Variations in the set of random parameters $\boldsymbol{\eta}_n^R$ are represented semi-parametrically with a discrete mixing distribution over a finite support set S. Consider the following logit-type expression for the probability that $\boldsymbol{\eta}_n^R = \boldsymbol{\eta}_r^R$:

$$w_n(\boldsymbol{\eta}_r^R|\boldsymbol{\alpha}) = \Pr(\boldsymbol{\eta}_n^R = \boldsymbol{\eta}_r^R) = \frac{e^{\mathbf{z}(\boldsymbol{\eta}_r^R)'\boldsymbol{\alpha}}}{\sum_{s \in S} e^{\mathbf{z}(\boldsymbol{\eta}_s^R)'\boldsymbol{\alpha}}} \tag{A.4}$$

where $\boldsymbol{\alpha}$ is a vector of parameters and $\mathbf{z}(\boldsymbol{\eta}_r^R)$ is a vector-valued function that captures the shape of the mixing distribution. $\mathbf{z}$ can be specified as a sieve function, such as polynomial or other functional forms, including step functions and splines (see details in Train, 2016).

The unconditional probability of the sequence of choices of individual n (P_n) is simply:

$$P_n(\boldsymbol{\delta}^F, \boldsymbol{\alpha}) = \sum_{r \in S} L_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R) w_n(\boldsymbol{\eta}_r^R|\boldsymbol{\alpha}), \tag{A.5}$$

where the parameters of interest are $\boldsymbol{\delta}^F$ and $\boldsymbol{\alpha}$.

A.2 Maximum Likelihood Estimator

Adopting a frequentist approach to the estimation of the parameters of interest, the maximum likelihood estimator is implemented. The loglikelihood of the LML-FR model is shown in equation 1.6:

$$\mathcal{L}(\boldsymbol{\delta}^F, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \ln\Big(\sum_{r \in S} L_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R) w_n(\boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) \Big). \tag{A.6}$$

The simulated loglikelihood can be then written as:

$$\tilde{\mathcal{L}}(\boldsymbol{\delta}^F, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \ln\Big(\sum_{r \in S_n} L_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R) w_n(\boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) \Big). \tag{A.7}$$

The partial derivative of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{\alpha}$ is:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\alpha}} = \sum_{n=1}^{N} \sum_{r \in S_n} \Big(h_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) - w_n(\boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) \Big) \mathbf{z}(\boldsymbol{\eta}_r^R) \tag{A.8}$$

where

$$h_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) = \frac{L_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R) w_n(\boldsymbol{\eta}_r^R | \boldsymbol{\alpha})}{\sum_{s \in S_n} L_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_s^R) w_n(\boldsymbol{\eta}_s^R | \boldsymbol{\alpha})}; \tag{A.9}$$

and the partial derivative of $\tilde{\mathcal{L}}$ with respect to $\boldsymbol{\delta}^F$ is:

$$\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\delta}^F} = \sum_{n=1}^{N} \sum_{r \in S_n} \left(\gamma_n^R h_n(\boldsymbol{\delta}^F, \boldsymbol{\eta}_r^R | \boldsymbol{\alpha}) \sum_{t=1}^{T} \left(\mathbf{x}_{ni_{nt}t}^F - \sum_{j=1}^{J} \mathbf{x}_{njt}^F Q_{njt} \right) \right). \tag{A.10}$$

Finally, the simulated score (gradient of $\tilde{\mathcal{L}}$) is:

$$\nabla(\tilde{\mathcal{L}}) = \left[\frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\alpha}} \quad \frac{\partial \tilde{\mathcal{L}}}{\partial \boldsymbol{\delta}^F} \right]. \tag{A.11}$$

APPENDIX B

APPENDIX OF CHAPTER 2

This section explores the lower-bound approximation of the weighted MNL Hessian. We first provide the sketch of the proof given by Böhning and Lindsay (1988) and then illustrate tightness of the approximation under different situations in a Monte Carlo study.

We used the lower-bound approximation of the Hessian of the weighted panel MNL model in the minorization-maximization algorithm, but to understand behavior of the approximation we consider the Hessian of an unweighted cross-section MNL model:

$$\boldsymbol{H} = -\sum_{i=1}^{N}\Big[\sum_{j=1}^{J}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}{}^{T}P_{ij}(\boldsymbol{\alpha}) - \Big(\sum_{j=1}^{J}\boldsymbol{x}_{ij}P_{ij}(\boldsymbol{\alpha})\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{ij}P_{ij}(\boldsymbol{\alpha})\Big)^{T}\Big],$$

where $\boldsymbol{x}_{ij}$ is the attribute vector of alternative j for person i, and $P_{ij}(\boldsymbol{\alpha})$ is the probability (conditional on $\boldsymbol{\alpha}$) of that person i choosing alternative j.

For an arbitrary $\boldsymbol{h}$, consider the alternative representation of the quadratic form (QF) of the Hessian:

$$QF1 = \boldsymbol{h}^{T}\left[-\sum_{i=1}^{N}\Big[\sum_{j=1}^{J}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}{}^{T}P_{ij}(\boldsymbol{\alpha}) - \Big(\sum_{j=1}^{J}\boldsymbol{x}_{ij}P_{ij}(\boldsymbol{\alpha})\Big)\Big(\sum_{j=1}^{J}\boldsymbol{x}_{ij}P_{ij}(\boldsymbol{\alpha})\Big)^{T}\Big]\right]\boldsymbol{h}$$

$$QF2 = -\sum_{i=1}^{N}\left[\sum_{j=1}^{J}(\boldsymbol{h}^{T}\boldsymbol{x}_{ij})^{2}P_{ij}(\boldsymbol{\alpha}) - \left(\sum_{j=1}^{J}(\boldsymbol{h}^{T}\boldsymbol{x}_{ij})P_{ij}(\boldsymbol{\alpha})\right)^{2}\right]$$

$$QF3 = -\sum_{i=1}^{N}\boldsymbol{h}^{T}\boldsymbol{x}_{i}\left(D(\boldsymbol{P}_{i}) - \boldsymbol{P}_{i}\boldsymbol{P}_{i}^{T}\right)\boldsymbol{x}_{i}^{T}\boldsymbol{h} = -\sum_{i=1}^{N}(\boldsymbol{h}^{T}\boldsymbol{x}_{i})M(\boldsymbol{P}_{i})(\boldsymbol{x}_{i}^{T}\boldsymbol{h}),$$

where $\boldsymbol{x}_i$ is a matrix of dimension $K \times J$ (K is the dimension of attributes and J is the number of alternatives), $\boldsymbol{P}_i$ is a column vector of choice probabilities for person i, and $D(\boldsymbol{P}_i)$ is a diagonal matrix with vector $\boldsymbol{P}_i$ in the diagonal.

Expression $QF3$ indicates that finding the global lower bound of the Hessian is the same as finding the global upper bound (which does not depend on the parameter $\boldsymbol{\alpha}$) of $M(\boldsymbol{P}_i)$. According to $QF2$, the problem is the same as maximizing the variance of a random variable ($\boldsymbol{h}^T \boldsymbol{x}_{ij}$) which has probability density function P_{ij}. Since the variance can be maximized by setting an equal weight of 0.5 on maximum and minimum values of ($\boldsymbol{h}^T \boldsymbol{x}_{ij}$), $\boldsymbol{P}_* = [.5, .5]$ is an optimal choice to obtain the lower bound of the Hessian.

The bound is sharp for two alternatives, and Böhning and Lindsay (1988) extended the bound for the multinomial case. The authors suggested to find a constant $c(J)$ (a constant as a function of J) such that $M(\boldsymbol{P}_i) \leq c(J) M(\boldsymbol{P}_*)$, where $\boldsymbol{P}_*$ is a vector of uniform multinomial probabilities (i.e., $[\frac{1}{J} \dots \frac{1}{J}]$). Using the same idea of weight allocation, Böhning and Lindsay (1988) found the value of $c(J) = \frac{J}{2}$ as optimal choice to obtain the lower bound of the Hessian. Since $\boldsymbol{h}$ can have any arbitrary value, we set it to the unit vector and insert the upper bound of $M(\boldsymbol{P}_i)$ in $QF3$ to obtain the approximation of Hessian ($B1$):

$$B1 = -\frac{1}{2} \sum_{i=1}^{N} \Big[\sum_{j=1}^{J} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \frac{1}{J} \Big(\sum_{j=1}^{J} \boldsymbol{x}_{ij} \Big) \Big(\sum_{j=1}^{J} \boldsymbol{x}_{ij} \Big)^T \Big].$$

Böhning and Lindsay (1988) noted that the magnitude of $\frac{1}{2} \sum_{i=1}^{N} \frac{1}{J} \Big(\sum_{j=1}^{J} \boldsymbol{x}_{ij} \Big) \Big(\sum_{j=1}^{J} \boldsymbol{x}_{ij} \Big)^T$, as compared to $-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{J} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T$, starts diminishing as the number of alterna-

tives grows. For a large choice set, $B1$ effectively becomes closer to $B2$:

$$B2 = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{J}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}{}^{T}.$$

To empirically understand the effect of the number of alternatives on the Hessian approximation, we considered several simulation scenarios with a unidimensional attribute ($K = 1$). The average of the approximate and true Hessian values across 50 repetitions are summarized in Table B.1. Each scenario is defined according to a combination of mean and standard deviations of the attribute, whether the actual PDF of $\boldsymbol{P}_i$ is uniform or not, and whether attributes are standardized or not. Consistent across all scenarios, the approximation $B1$ performs well for 2 alternatives. In fact, it coincides with the true value of the Hessian when the actual distribution of the choice probabilities is uniform. As expected, $B1$ is slightly away from the true Hessian even for the two-alternative case when choice probabilities are simulated randomly. $B1$ becomes a poorer lower bound approximation of the Hessian with an increase in the number of alternatives, which can be attributed to a poor extension of the binary analogy to the multinomial case – there is no strong reason to use $M(\boldsymbol{P}_i) \leq c(J)M(\boldsymbol{P}_*)$. As the number of alternatives increases, $B1 - B2$ increases but the relative contribution of $B1 - B2$ decreases and thus $B1$ attains a value closer to $B2$. However, comparison of scenarios 3, 4, and 5 indicates that the percentage difference between $B1$ and the true Hessian is not affected by the change in standard deviation of the attributes, but $B2$ does become worse when the standard deviation is lower. This observation supports the original findings of Böhning and Lindsay (1988) who noted that $B1$ is unaffected by standardization of the attributes, but $B2$ lacks this feature.

Table B.1: Lower Bound Approximation of Hessian: Simulation Results

Scenario	J	Mean	Std. Dev.	PDF[a]	Standardized[b]	B1	B2	B1-B2	B1/B1-B2	H	(H-B1)/H*100	(H-B2)/H*100	(H-B2)/(H-B1)
1	2	0	1	Uniform	No	-501	-1000	499	-1.0	-501.2	0	-99.5	∞
1	5	0	1	Uniform	No	-1997	-2498	501	-4.0	-798.7	-150	-212.7	1.4
1	10	0	1	Uniform	No	-4509	-5011	502	-9.0	-901.7	-400	-455.7	1.1
1	100	0	1	Uniform	No	-49470	-49967	496	-99.9	-989.4	-4900	-4950.2	1.0
2	2	0	1	Random	No	-500	-1002	502	-1.0	-395.8	-26	-153.3	5.8
2	5	0	1	Random	No	-2002	-2502	501	-4.0	-733.8	-173	-241.0	1.4
2	10	0	1	Random	No	-4475	-4975	500	-9.0	-862.9	-419	-476.6	1.1
2	100	0	1	Random	No	-49535	-50039	504	-98.4	-987.4	-4917	-4967.6	1.0
3	2	2	2	Uniform	No	-2012	-7988	5976	-0.3	-2012.1	0	-297.0	∞
3	5	2	2	Uniform	No	-7993	-20071	12077	-0.7	-3197.4	-150	-527.7	3.5
3	10	2	2	Uniform	No	-18025	-40059	22034	-0.8	-3605.0	-400	-1011.2	2.5
3	100	2	2	Uniform	No	-198234	-400504	202270	-1.0	-3964.7	-4900	-10001.8	2.0
4	2	2	0.5	Uniform	No	-125	-4255	4130	-0.030	-125.1	0	-3301.7	∞
4	5	2	0.5	Uniform	No	-501	-10633	10132	-0.049	-200.5	-150	-5203.3	34.7
4	10	2	0.5	Uniform	No	-1125	-21257	20132	-0.056	-225.0	-400	-9346.1	23.4
4	100	2	0.5	Uniform	No	-12365	-212482	200117	-0.062	-247.3	-4900	-85819.5	17.5
5	2	2	0.1	Uniform	No	-5	-4009	4004	-0.0012	-5.0	0	-80223.5	∞
5	5	2	0.1	Uniform	No	-20	-10024	10004	-0.0020	-8.0	-150	-125232.4	834.9
5	10	2	0.1	Uniform	No	-45	-20050	20005	-0.0022	-9.0	-400	-222740.0	556.9
5	100	2	0.1	Uniform	No	-495	-200509	200014	-0.0025	-9.9	-4900	-2024817.9	413.2
6	2	2	2	Uniform	Yes	-503	-1008	505	-1.0	-502.7	0	-100.5	∞
6	5	2	2	Uniform	Yes	-2007	-2507	500	-4.0	-802.7	-150	-212.2	1.4
6	10	2	2	Uniform	Yes	-4494	-4998	504	-8.9	-898.8	-400	-456.1	1.1
6	100	2	2	Uniform	Yes	-49484	-49984	500	-99.2	-989.7	-4900	-4950.5	1.0

[a] Probability density function across alternatives; uniform $\boldsymbol{P}_i$: $[\frac{1}{J} \dots \frac{1}{J}]$.

[b] Standardization - subtract mean from the attribute and divide by standard deviation.

APPENDIX C

APPENDIX OF CHAPTER 4

C.1 Matrix transformations

C.1.1 Transformation matrix (D) to compute $\boldsymbol{\Lambda}$ from $\overline{\boldsymbol{\Lambda}}$

We show below the relation between the variance-covariance matrix of the error ($\boldsymbol{\Lambda}_{I_K \times I_K}$) and the normalized variance-covariance matrix of the error difference ($\overline{\boldsymbol{\Lambda}}_{(I_K-1)\times(I_K-1)}$):

$$\overline{\boldsymbol{\Lambda}} = \begin{bmatrix} \overline{\boldsymbol{\Lambda}}_1 & \overline{\boldsymbol{\Lambda}}_{1,2} & \dots & \overline{\boldsymbol{\Lambda}}_{1,I-1} \\ \overline{\boldsymbol{\Lambda}}_{2,1} & \overline{\boldsymbol{\Lambda}}_2 & \dots & \overline{\boldsymbol{\Lambda}}_{2,I-1} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\boldsymbol{\Lambda}}_{I-1,1} & \overline{\boldsymbol{\Lambda}}_{I-1,2} & \dots & \overline{\boldsymbol{\Lambda}}_{I-1} \end{bmatrix}_{(I_K-I)\times(I_K-I)}$$

$$\text{where} \quad \overline{\boldsymbol{\Lambda}}_i = \begin{bmatrix} 1 & \overline{\Lambda}^i_{1,2} & \dots & \overline{\Lambda}^i_{1,i_K-1} \\ \overline{\Lambda}^i_{2,1} & \overline{\Lambda}^i_{2,2} & \dots & \overline{\Lambda}^i_{2,i_K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\Lambda}^i_{i_K-1,1} & \overline{\Lambda}^i_{i_K-1,2} & \dots & \overline{\Lambda}^i_{i_K-1,i_K-1} \end{bmatrix}_{(i_K-1)\times(i_K-1)} \tag{C.1}$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_{1,2} & \dots & \boldsymbol{\Lambda}_{1,I} \\ \boldsymbol{\Lambda}_{2,1} & \boldsymbol{\Lambda}_2 & \dots & \boldsymbol{\Lambda}_{2,I} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda}_{I,1} & \boldsymbol{\Lambda}_{I,2} & \dots & \boldsymbol{\Lambda}_I \end{bmatrix}_{(I_K)\times(I_K)} \tag{C.2}$$

$$\text{where } \boldsymbol{\Lambda}_i = \begin{bmatrix} 0 & \mathbf{0}_{1,i_K-1} \\ \mathbf{0}_{i_K-1,1} & \overline{\boldsymbol{\Lambda}}_i \end{bmatrix}_{i_K \times i_K} \text{ and } \boldsymbol{\Lambda}_{i,j} = \begin{bmatrix} 0 & \mathbf{0}_{1,j_K-1} \\ \mathbf{0}_{i_K-1,1} & \overline{\boldsymbol{\Lambda}}_{i,j} \end{bmatrix}_{i_K \times j_K}$$

Note that $\boldsymbol{\Lambda} = \boldsymbol{D}\overline{\boldsymbol{\Lambda}}\boldsymbol{D}^{\intercal}$, where the transformation matrix $\boldsymbol{D}$ is constructed based on algorithm 5.

Algorithm 5: Creating $\boldsymbol{D}$ matrix

Initialization: $\boldsymbol{D} = \mathbf{0}_{I_K \times (I_K - 1)}$
for *(m in 1 to I)* **do**
  if $m == 1$ **then**
    $R_1 = 2$;
    $R_2 = m_K$;
    $C_1 = 1$;
    $C_2 = m_K - 1$;
  else
    $R_1 = \sum_{n=1}^{m-1} n_K + 2$;
    $R_2 = \sum_{n=1}^{m} n_K$;
    $C_1 = \sum_{n=1}^{m-1} (n_K - 1) + 1$;
    $C_2 = \sum_{n=1}^{m} (n_K - 1)$;
  end
  $\boldsymbol{D}(R_1 : R_2, C_1 : C_2) = \mathbf{1}_{(m_K - 1) \times (m_K - 1)}$;
end
Note: $\mathbf{1}_{i \times i}$ and $\mathbf{0}_{i \times i}$ are identity matrix and matrix of zeros, respectively, of size $i \times i$.

We provide an example to illustrate the transformation from $\overline{\boldsymbol{\Lambda}}$ to $\boldsymbol{\Lambda}$ using $\boldsymbol{D}$. We consider a case of $I = 2$, with $1_K = 3$ and $2_K = 4$. The $\overline{\boldsymbol{\Lambda}}_{5\times5}$, $\boldsymbol{\Lambda}_{7\times7}$, and the transformation matrix $\boldsymbol{D}_{7\times5}$ would be:

$$\overline{\mathbf{\Lambda}} = \begin{bmatrix} \overline{\mathbf{\Lambda}}_1 & \overline{\mathbf{\Lambda}}_{1,2} \\ \overline{\mathbf{\Lambda}}_{2,1} & \overline{\mathbf{\Lambda}}_2 \end{bmatrix} = \left[\begin{array}{ccc|cc} 1 & \overline{\Lambda}^{1}_{1,2} & \overline{\Lambda}^{1,2}_{1,1} & \overline{\Lambda}^{1,2}_{1,2} & \overline{\Lambda}^{1,2}_{1,3} \\ \overline{\Lambda}^{1}_{2,1} & \overline{\Lambda}^{1}_{2,2} & \overline{\Lambda}^{1,2}_{2,1} & \overline{\Lambda}^{1,2}_{2,2} & \overline{\Lambda}^{1,2}_{2,3} \\ \hline \overline{\Lambda}^{2,1}_{1,1} & \overline{\Lambda}^{2,1}_{1,2} & 1 & \overline{\Lambda}^{2}_{1,2} & \overline{\Lambda}^{2}_{1,3} \\ \overline{\Lambda}^{2,1}_{2,1} & \overline{\Lambda}^{2,1}_{2,2} & \overline{\Lambda}^{2}_{2,1} & \overline{\Lambda}^{2}_{2,2} & \overline{\Lambda}^{2}_{2,3} \\ \overline{\Lambda}^{2,1}_{3,1} & \overline{\Lambda}^{2,1}_{3,2} & \overline{\Lambda}^{2}_{3,1} & \overline{\Lambda}^{2}_{3,2} & \overline{\Lambda}^{2}_{3,3} \end{array}\right]$$

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{\Lambda}_{1,2} \\ \mathbf{\Lambda}_{2,1} & \mathbf{\Lambda}_2 \end{bmatrix} == \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \overline{\Lambda}^{1}_{1,2} & 0 & \overline{\Lambda}^{1,2}_{1,1} & 0\overline{\Lambda}^{1,2}_{1,2} & \overline{\Lambda}^{1,2}_{1,3} \\ 0 & \overline{\Lambda}^{1}_{2,1} & \overline{\Lambda}^{1}_{2,2} & 0 & \overline{\Lambda}^{1,2}_{2,1} & \overline{\Lambda}^{1,2}_{2,2} & \overline{\Lambda}^{1,2}_{2,3} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \overline{\Lambda}^{2,1}_{1,1} & \overline{\Lambda}^{2,1}_{1,2} & 0 & 1 & \overline{\Lambda}^{2}_{1,2} & \overline{\Lambda}^{2}_{1,3} \\ 0 & \overline{\Lambda}^{2,1}_{2,1} & \overline{\Lambda}^{2,1}_{2,2} & 0 & \overline{\Lambda}^{2}_{2,1} & \overline{\Lambda}^{2}_{2,2} & \overline{\Lambda}^{2}_{2,3} \\ 0 & \overline{\Lambda}^{2,1}_{3,1} & \overline{\Lambda}^{2,1}_{3,2} & 0 & \overline{\Lambda}^{2}_{3,1} & \overline{\Lambda}^{2}_{3,2} & \overline{\Lambda}^{2}_{3,3} \end{array}\right] \qquad \text{(C.3)}$$

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

C.1.2 Modified transformation matrix ($\boldsymbol{D}_m$) to compute $\boldsymbol{\Sigma}$ from $\overline{\boldsymbol{\Sigma}}$

The *modified transformation matrix* ($\boldsymbol{D}_m$) can be easily computed from the *transformation matrix* ($\boldsymbol{D}$) by appending an identity matrix as follows:

$$\boldsymbol{D}_m = \begin{bmatrix} \mathbf{1}_{H\times H} & \mathbf{0}_{H\times(I_K-I)} \\ \mathbf{0}_{I_K\times H} & \boldsymbol{D} \end{bmatrix}_{(H+I_K)\times(H+I_K-I)} \tag{C.4}$$

where H is the number of continuous outcomes, $\mathbf{1}_{H\times H}$ is an identity matrix of size $H\times H$, and $\boldsymbol{D}$ is obtained from algorithm 5. We expand on the example of appendix C.1.1 which considers $I = 2$ with $1_K = 3$ and $2_K = 4$. If $H = 2$, then the *modified transformation matrix* $\boldsymbol{D}_m$ would be:

$$\begin{aligned} \boldsymbol{D}_m &= \begin{bmatrix} \mathbf{1}_{2\times2} & \mathbf{0}_{2\times5} \\ \mathbf{0}_{7\times2} & \boldsymbol{D}_{7\times5} \end{bmatrix} \\ &= \left[\begin{array}{cc|ccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{aligned} \tag{C.5}$$

C.1.3 Utility difference generator (M) to compute $\widetilde{\Sigma}$ from Σ

Algorithm 6: Creating $\boldsymbol{M}$ matrix

Initialisation: $\boldsymbol{M} = \mathbf{0}_{(H+I_K-I)\times(H+I_K)}$;
$\boldsymbol{M}(1:H, 1:H) = \mathbf{1}_{H\times H}$;
for *(n in 1 to I)* **do**
  $F_1 = \mathbf{1}_{(n_K-1)\times(n_K-1)}$;
  $T_1 = -\boldsymbol{O}_{n_K-1}$;
  if $n_m == 1$ **then**
    $S_1 = T_1 \sim F_1$;
  else
    if $n_m == n_K$ **then**
      $S_1 = F_1 \sim T_1$;
    else
      $S_1 = F_1[, 1:(n_m-1)] \sim T_1 \sim F_1[, n_m:(n_K-1)]$;
    end
  end
  if $n == 1$ **then**
    $R_1 = H + 1$;
    $R_2 = H + n_K - 1$;
    $C_1 = H + 1$;
    $C_2 = H + n_K$;
  else
    $R_1 = H + \left(\sum_{j=1}^{n-1}(j_K - 1)\right) + 1$;
    $R_2 = H + \left(\sum_{j=1}^{n}(j_K - 1)\right)$;
    $C_1 = H + \left(\sum_{j=1}^{n-1} j_K\right) + 1$;
    $C_2 = H + \left(\sum_{j=1}^{n} j_K\right)$;
  end
  $\boldsymbol{M}(R_1:R_2, C_1:C_2) = S_1$;
end

Note 1: $\mathbf{1}_{i\times i}$ and $\mathbf{0}_{i\times i}$ are identity matrix and matrix of zeros, respectively, of size $i \times i$.
Note 2: $\boldsymbol{O}_i$ is a column vector of ones of size $i \times 1$. " $\sim$ " implies horizontal concatenation.
Note 3: n_m is the index of the chosen alternative for the nominal variable n.

We consider the same example of appendix C.1.2 which considers $H = 2$ and $I = 2$ with $1_K = 3$, $1_m = 2$, $2_K = 4$, and $2_m = 3$. We use algorithm 6 to construct $\boldsymbol{M}$ matrix for this example:

$$\boldsymbol{M} = \left[\begin{array}{cc|ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{array}\right] \tag{C.6}$$

C.1.4 Reparametrization of the Cholesky decomposition of $\overline{\boldsymbol{\Sigma}}$

Consider $\boldsymbol{L}\boldsymbol{L}^\intercal = \overline{\boldsymbol{\Sigma}}$, where $\boldsymbol{L}$ is the lower triangular Cholesky matrix of size $(H + I_K - I) \times (H + I_K - I)$. We reparametrize all rows of the $\boldsymbol{L}$ matrix for which the diagonal element of $\overline{\boldsymbol{\Sigma}}$ is normalized to 1. We compute $a_i = \sqrt{1 + \sum_{j=1}^{i-1} L_{i,j}^2}$ for the i^{th} row and modify non-diagonal elements $L_{i,r} = \frac{L_{i,r}}{a_i} \forall r \in \{1, 2, \ldots, i-1\}$ and the diagonal element $L_{i,i} = \frac{1}{a_i}$.

C.2 MVTNCD illustration

The derivation of equation 4.8 from equation 4.6 can be understood easily based on a transformation applied on a bivariate t-cumulative distribution function. For $p = 2$, equation 4.6 after substitution can be written as (note that $w_1 = u_1$):

$$T_p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Omega}, \delta) = \kappa_\delta^2 \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+2)}{2}} \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1 \left(du_2 \sqrt{\frac{\delta + u_1^2}{\delta + 1}}\right)$$

$$= \kappa_\delta^2 \left[\int_{\hat{a}_1}^{\hat{b}_1} \sqrt{\frac{\delta + u_1^2}{\delta + 1}} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+2)}{2}} du_1\right] \left[\int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_2\right]$$

$$= \kappa_\delta^2 \left[\int_{\hat{a}_1}^{\hat{b}_1} \sqrt{\left(\frac{\delta}{\delta + 1}\right)\left(1 + \frac{u_1^2}{\delta}\right)} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+2)}{2}} du_1\right] \left[\int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_2\right]$$

$$= \kappa_\delta^2 \sqrt{\frac{\delta}{\delta + 1}} \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1 du_2 \quad \text{(equation 4.7)}$$

Now deriving equation 4.8

$$= \frac{\Gamma\left(\frac{\delta+2}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)(\pi\delta)} \sqrt{\frac{\delta}{\delta + 1}} \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1 du_2$$

$$= \frac{\Gamma\left(\frac{\delta+1}{2}\right)\Gamma\left(\frac{\delta+1+1}{2}\right)}{\Gamma\left(\frac{\delta}{2}\right)(\pi\delta)^{\frac{1}{2}}\Gamma\left(\frac{\delta+1}{2}\right)(\pi(\delta + 1))^{\frac{1}{2}}} \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1 du_2$$

$$= \kappa_\delta^1 \kappa_{\delta+1}^1 \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1 du_2$$

$$= \left[\kappa_{\delta+1-1}^1 \int_{\hat{a}_1}^{\hat{b}_1} \left(1 + \frac{u_1^2}{\delta}\right)^{-\frac{(\delta+1)}{2}} du_1\right] \left[\kappa_{\delta+2-1}^1 \int_{\hat{a}_2}^{\hat{b}_2} \left(1 + \frac{u_2^2}{\delta + 2 - 1}\right)^{-\frac{(\delta+2)}{2}} du_1\right] \text{(equation 4.8)}$$

BIBLIOGRAPHY

Abay, K. A. (2015), 'Evaluating simulation-based approaches and multivariate quadrature on sparse grids in estimating multivariate binary probit models', *Economics Letters* **126**, 51–56.

Achtnicht, M., Bühler, G. and Hermeling, C. (2012), 'The impact of fuel availability on demand for alternative-fuel vehicles', *Transportation Research Part D: Transport and Environment* **17**(3), 262–269.

Ahsanullah, M., Kibria, B. G. and Shakil, M. (2014), Normal distribution, *in* 'Normal and Student´s t Distributions and Their Applications', Springer, pp. 7–50.

Askey, R. (1975), *Orthogonal polynomials and special functions*, Vol. 21, Siam.

Azzalini, A. and Arellano-Valle, R. B. (2013), 'Maximum penalized likelihood estimation for skew-normal and skew-t distributions', *Journal of Statistical Planning and Inference* **143**(2), 419–433.

Azzalini, A. and Genton, M. G. (2008), 'Robust likelihood methods based on the skew-t and related distributions', *International Statistical Review* **76**(1), 106–129.

Bajari, P., Fox, J. T. and Ryan, S. P. (2007), 'Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients', *The American Economic Review* **97**(2), 459–463.

Bansal, P., Daziano, R. A. and Achtnicht, M. (2018*a*), 'Comparison of parametric and semiparametric representations of unobserved preference heterogeneity in logit models', *Journal of Choice Modelling* **27**, 97–113.

Bansal, P., Daziano, R. A. and Achtnicht, M. (2018*b*), 'Extending the logit-mixed logit model for a combination of random and fixed parameters', *Journal of Choice Modelling* **27**, 88–96.

Bansal, P., Daziano, R. A. and Achtnicht, M. (2018*c*), 'Extending the logit-mixed logit model for a combination of random and fixed parameters', *Journal of choice modelling* **27**, 88–96.

Bastin, F., Cirillo, C. and Toint, P. L. (2010), 'Estimating nonparametric random utility models with an application to the value of time in heterogeneous populations', *Transportation Science* **44**(4), 537–549.

Bazán, J. L., Bolfarine, H. and Branco, M. D. (2010), 'A framework for skew-probit links in binary regression', *Communications in Statistics—Theory and Methods* **39**(4), 678–697.

Beck, M. J., Rose, J. M. and Hensher, D. A. (2013), 'Consistently inconsistent: The role of certainty, acceptability and scale in choice', *Transportation Research Part E: Logistics and Transportation Review* **56**, 81–93.

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T. and Polydoropoulou, A. (2002), 'Integration of choice and latent variable models', *Perpetual motion: Travel behaviour research opportunities and application challenges* pp. 431–470.

Bhaduri, A. and Graham-Brady, L. (2018), 'An efficient adaptive sparse grid collo-

cation method through derivative estimation', *Probabilistic Engineering Mechanics* **51**, 11–22.

Bhat, C. R. (1997), 'An endogenous segmentation mode choice model with an application to intercity travel', *Transportation Science* **31**(1), 34–48.

Bhat, C. R. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model', *Transportation Research Part B: Methodological* **35**(7), 677–693.

Bhat, C. R. (2003), 'Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences', *Transportation Research Part B: Methodological* **37**(9), 837–855.

Bhat, C. R. (2011), 'The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models', *Transportation Research Part B: Methodological* **45**(7), 923–939.

Bhat, C. R. (2014), 'The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models', *Foundations and Trends® in Econometrics* **7**(1), 1–117.

Bhat, C. R. (2015), 'A new generalized heterogeneous data model (ghdm) to jointly model mixed types of dependent variables', *Transportation Research Part B: Methodological* **79**, 50–77.

Bhat, C. R., Astroza, S. and Hamdi, A. S. (2017), 'A spatial generalized ordered-

response model with skew normal kernel error terms with an application to bicycling frequency', *Transportation Research Part B: Methodological* **95**, 126–148.

Bhat, C. R., Dubey, S. K. and Nagel, K. (2015), 'Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice', *Transportation Research Part B: Methodological* **78**, 341–363.

Bhat, C. R. and 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.

Bhat, C. R. and Lavieri, P. S. (2017), 'A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions', *Theory and Decision* pp. 1–37.

Bhat, C. R. and Sidharthan, R. (2012), 'A new approach to specify and estimate non-normally mixed multinomial probit models', *Transportation Research Part B: Methodological* **46**(7), 817–833.

Bierlaire, M. (2016), PythonBiogeme: a short introduction, Technical report, TRANSP-OR 160706. Transport and Mobility Laboratory, ENAC, EPFL.

Böhning, D. and Lindsay, B. G. (1988), 'Monotonicity of quadratic-approximation algorithms', *Annals of the Institute of Statistical Mathematics* **40**(4), 641–663.

Börger, T. (2016), 'Are fast responses more random? testing the effect of response time on scale in an online choice experiment', *Environmental and Resource Economics* **65**(2), 389–413.

Boyd, J. H. and Mellman, R. E. (1980), 'The effect of fuel economy standards on the us automotive market: an hedonic demand analysis', *Transportation Research Part A: General* **14**(5-6), 367–378.

Brathwaite, T. and Walker, J. L. (2018), 'Asymmetric, closed-form, finite-parameter models of multinomial choice', *Journal of Choice Modelling* **29**, 78–112.

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

Brumm, J. and Scheidegger, S. (2017), 'Using adaptive sparse grids to solve high-dimensional dynamic models', *Econometrica* **85**(5), 1575–1612.

Bunch, D. S., Bradley, M., Golob, T. F., Kitamura, R. and Occhiuzzo, G. P. (1993), 'Demand for clean-fuel vehicles in california: a discrete-choice stated preference pilot project', *Transportation Research Part A: Policy and Practice* **27**(3), 237–253.

Cagnone, S. and Bartolucci, F. (2017), 'Adaptive quadrature for maximum likelihood estimation of a class of dynamic latent variable models', *Computational Economics* **49**(4), 599–622.

Camilleri, L. (2009), 'Bias of standard errors in latent class model applications using Newton-Raphson and EM algorithms.', *Journal of Advanced Computational Intelligence and Intelligent Informatics* **13**(5), 537–541.

Castillo, E., Menéndez, J. M., Jiménez, P. and Rivas, A. (2008), 'Closed form expressions for choice probabilities in the weibull case', *Transportation Research Part B: Methodological* **42**(4), 373–380.

Cherchi, E. and Guevara, C. A. (2012), 'A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance–covariance matrix', *Transportation Research Part B: Methodological* **46**(2), 321–332.

Cherchi, E. and Polak, J. W. (2005), 'Assessing user benefits with discrete choice models: Implications of specification errors under random taste heterogeneity', *Transportation Research Record* **1926**(1), 61–69.

Cherry, C. and Cervero, R. (2007), 'Use characteristics and mode choice behavior of electric bike users in China', *Transport Policy* **14**(3), 247–257.

Clark, W. A., Huang, Y. and Withers, S. (2003), 'Does commuting distance matter?: Commuting tolerance and residential change', *Regional Science and Urban Economics* **33**(2), 199–221.

CMC (2017), CMC choice modelling code for R, Technical report, Choice Modelling Centre, University of Leeds, www.cmc.leeds.ac.uk.

Craig, P. (2008), 'A new reconstruction of multivariate normal orthant probabilities', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 227–243.

Davis, P. J. and Rabinowitz, P. (2007), *Methods of numerical integration*, Courier Corporation.

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

De Leon, A. R. and Chough, K. C. (2013), *Analysis of mixed data: methods & applications*, CRC Press.

Dekker, T., Hess, S., Brouwer, R. and Hofkes, M. (2016), 'Decision uncertainty in multi-attribute stated preference studies', *Resource and Energy Economics* **43**, 57–73.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society Series B (Methodological)* pp. 1–38.

DeSarbo, W. S., Ramaswamy, V. and Cohen, S. H. (1995), 'Market segmentation with choice-based conjoint analysis', *Marketing Letters* **6**(2), 137–147.

Dick, J. and Pillichshammer, F. (2014), Discrepancy theory and quasi-monte carlo integration, *in* 'A Panorama of Discrepancy Theory', Springer, pp. 539–619.

Ding, P. (2016), 'On the conditional distribution of the multivariate t distribution', *The American Statistician* **70**(3), 293–295.

Elrod, T., Johnson, R. D. and White, J. (2004), 'A new integrated model of non-compensatory and compensatory decision strategies', *Organizational Behavior and Human Decision Processes* **95**(1), 1–19.

Ewing, R. and Cervero, R. (2010), 'Travel and the built environment: a meta-analysis', *Journal of the American Planning Association* **76**(3), 265–294.
URL: *http://www.informaworld.com/10.1080/01944361003766766*

Fang, H. A. (2008), 'A discrete–continuous model of households' vehicle choice and usage, with an application to the effects of residential density', *Transportation Research Part B: Methodological* **42**(9), 736–758.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* pp. 209–230.

Fisher, C., Bashyal, S. and Bachman, B. (2012), 'Demographic impacts on environmentally friendly purchase behaviors', *Journal of Targeting, Measurement and Analysis for Marketing* **20**(3-4), 172–184.

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

Fosgerau, M. and Bierlaire, M. (2007), 'A practical test for the choice of mixing distribution in discrete choice models', *Transportation Research Part B: Methodological* **41**(7), 784–794.

Fosgerau, M. and Bierlaire, M. (2009), 'Discrete choice models with multiplicative error terms', *Transportation Research Part B: Methodological* **43**(5), 494–505.

Fosgerau, M. and Hess, S. (2007), Competing methods for representing random taste heterogeneity in discrete choice models, Technical report, Working paper, Danish Transport Research Institute, Copenhagen.

Fosgerau, M. and Hess, S. (2009), 'A comparison of methods for representing random taste heterogeneity in discrete choice models', *European Transport-Trasporti Europei* **42**, 1–25.

Fosgerau, M. and Mabit, S. L. (2013), 'Easy and flexible mixture distributions', *Economics Letters* **120**(2), 206–210.

Fox, J. T., Ryan, S. P. and Bajari, P. (2011), 'A simple estimator for the distribution of random coefficients', *Quantitative Economics* **2**(3), 381–418.

Franceschinis, C., Scarpa, R. and Thiene, M. (2017), A Monte Carlo evaluation of the logit-mixed logit under asymmetry and multimodality, Technical report, Working Paper, University of Waikato, Hamilton.

Genz, A. (1992), 'Numerical computation of multivariate normal probabilities', *Journal of computational and graphical statistics* **1**(2), 141–149.

Genz, A. and Bretz, F. (1999), 'Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts', *Journal of Statistical Computation and Simulation* **63**(4), 103–117.

Geweke, J., Keane, M. and Runkle, D. (1994), 'Alternative computational approaches to inference in the multinomial probit model', *The review of economics and statistics* pp. 609–632.

Golub, G. H. and Welsch, J. H. (1969), 'Calculation of gauss quadrature rules', *Mathematics of computation* **23**(106), 221–230.

Goos, P. and Mylona, K. (2018), 'Quadrature methods for bayesian optimal design of experiments with nonnormal prior distributions', *Journal of Computational and Graphical Statistics* **27**(1), 179–194.

Greene, W. H. and Hensher, D. A. (2013), 'Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model', *Applied Economics* **45**(14), 1897–1902.

Guerra, E. (2017), 'Electric vehicles, air pollution, and the motorcycle city: A stated preference survey of consumers' willingness to adopt electric motorcycles in Solo, Indonesia', *Transportation Research Part D: Transport and Environment* .

Guevara, C. A., Cherchi, E. and Moreno, M. (2009), 'Estimating random coefficient logit models with full covariance matrix: comparing performance of mixed logit and laplace approximation methods', *Transportation Research Record* **2132**(1), 87–94.

Hajivassiliou, V., McFadden, D. and Ruud, P. (1996), 'Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results', *Journal of econometrics* **72**(1-2), 85–134.

Heiss, F. (2010), The panel probit model: Adaptive integration on sparse grids, *in* 'Maximum simulated likelihood methods and applications', Emerald Group Publishing Limited, pp. 41–64.

Heiss, F. and Winschel, V. (2008), 'Likelihood approximation by numerical integration on sparse grids', *journal of Econometrics* **144**(1), 62–80.

Hensher, D. A. and Greene, W. H. (2003), 'The mixed logit model: the state of practice', *Transportation* **30**(2), 133–176.

Hess, S., Train, K. E. and Polak, J. W. (2006), 'On the use of a modified latin hypercube sampling (mlhs) method in the estimation of a mixed logit model for vehicle choice', *Transportation Research Part B: Methodological* **40**(2), 147–163.

Jakeman, J. D. and Narayan, A. (2018), 'Generation and application of multivariate polynomial quadrature rules', *Computer Methods in Applied Mechanics and Engineering* **338**, 134–161.

Jakobsson, N., Gnann, T., Plötz, P., Sprei, F. and Karlsson, S. (2016), 'Are multi-car households better suited for battery electric vehicles?–driving patterns and economics in sweden and germany', *Transportation Research Part C: Emerging Technologies* **65**, 1–15.

James, J. (2017), 'MM algorithm for general mixed multinomial logit models', *Journal of Applied Econometrics* **32**(4), 841–857.

Jamshidian, M. and Jennrich, R. I. (2000), 'Standard errors for EM estimation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 257–270.

Jones, L. R., Cherry, C. R., Vu, T. A. and Nguyen, Q. N. (2013), 'The effect of incentives and technology on the adoption of electric motorcycles: A stated choice experiment in Vietnam', *Transportation Research Part A: Policy and Practice* **57**, 1–11.

Jones, M. (2002), 'A dependent bivariate t distribution with marginals on different degrees of freedom', *Statistics & probability letters* **56**(2), 163–170.

Kamakura, W. A. and Russell, G. J. (1989), 'A probabilistic choice model for market segmentation and elasticity structure', *Journal of marketing research* **26**(4), 379–390.

Keane, M. and Wasi, N. (2013), 'Comparing alternative models of heterogeneity in consumer choice behavior', *Journal of Applied Econometrics* **28**(6), 1018–1045.

Keshavarzzadeh, V., Kirby, R. M. and Narayan, A. (2018), 'Numerical integration in multiple dimensions with designed quadrature', *SIAM Journal on Scientific Computing* **40**(4), A2033–A2061.

Kim, S., Chen, M.-H. and Dey, D. K. (2007), 'Flexible generalized t-link models for binary response data', *Biometrika* **95**(1), 93–106.

Klein, N. J., Guerra, E. and Smart, M. J. (2018), 'The Philadelphia story: Age, race, gender and changing travel trends', *Journal of Transport Geography* **69**, 19–25. **URL:** *http://www.sciencedirect.com/science/article/pii/S0966692317307044*

Koehler, E., Brown, E. and Haneuse, S. J.-P. (2009), 'On the assessment of monte carlo error in simulation-based statistical analyses', *The American Statistician* **63**(2), 155–162.

Koenker, R. and Yoon, J. (2009), 'Parametric links for binary choice models: A fisherian–bayesian colloquy', *Journal of Econometrics* **152**(2), 120–130.

Kuhfeld, W. F., Tobias, R. D. and Garratt, M. (1994), 'Efficient experimental design with marketing research applications', *Journal of Marketing Research* pp. 545–557.

Lange, K., Hunter, D. R. and Yang, I. (2000), 'Optimization transfer using surrogate objective functions', *Journal of Computational and Graphical Statistics* **9**(1), 1–20.

Lee, B. S. and McDonald, J. F. (2003), 'Determinants of commuting time and distance for seoul residents: The impact of family status on the commuting of women', *Urban Studies* **40**(7), 1283–1302.

Li, B. (2011), 'The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis', *Transportation Research Part B: Methodological* **45**(3), 461–473.

Liu, C. (2004), 'Robit regression: A simple robust alternative to logistic and probit regression', *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* pp. 227–238.

Liu, Y., Bansal, P., Daziano, R. and Samaranayake, S. (2018), 'A framework to integrate mode choice in the design of mobility-on-demand systems', *Transportation Research Part C: Emerging Technologies* .

Louviere, J. and Eagle, T. (2006), Confound it! that pesky little scale constant messes up our convenient assumptions, *in* 'Proceedings of the Sawtooth Software Conference', pp. 211–228.

Louviere, J. J. and Meyer, R. J. (2008), 'Formal choice models of informal choices'.

Lundhede, T. H., Olsen, S. B., Jacobsen, J. B. and Thorsen, B. J. (2009), 'Handling respondent uncertainty in choice experiments: evaluating recoding approaches against explicit modelling of uncertainty', *Journal of Choice Modelling* **2**(2), 118–147.

Ma, X. and Zabaras, N. (2009), 'An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations', *Journal of Computational Physics* **228**(8), 3084–3113.

Marchenko, Y. V. and Genton, M. G. (2012), 'A heckman selection-t model', *Journal of the American Statistical Association* **107**(497), 304–317.

Martínez, F., Aguila, F. and Hurtubia, R. (2009), 'The constrained multinomial logit: A semi-compensatory choice model', *Transportation Research Part B: Methodological* **43**(3), 365–377.

McFadden, D. (1973), 'Conditional logit analysis of qualitative choice behavior', *Frontiers in Econometrics* pp. 105–142.

McFadden, D. and Train, K. (2000), 'Mixed MNL models for discrete response', *Journal of Applied Econometrics* pp. 447–470.

McLachlan, G. and Krishnan, T. (2007), *The EM algorithm and Extensions*, Vol. 382, second edn, John Wiley & Sons, Hoboken, New Jersey.

McQuaid, R. W. and Chen, T. (2012), 'Commuting times–the role of gender, children and part-time work', *Research in transportation economics* **34**(1), 66–73.

Meilijson, I. (1989), 'A fast improvement to the EM algorithm on its own terms', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 127–138.

Meng, X.-L. and Rubin, D. B. (1991), 'Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm', *Journal of the American Statistical Association* **86**(416), 899–909.

Muehlegger, E. and Rapson, D. (2018), Understanding the distributional impacts of vehicle policy: Who buys new and used alternative vehicles?, Technical report, National Center for Sustainable Transportation.

Munger, D., L'Ecuyer, P., Bastin, F., Cirillo, C. and Tuffin, B. (2012), 'Estimation of the mixed logit likelihood function by randomized quasi-monte carlo', *Transportation Research Part B: Methodological* **46**(2), 305–320.

Nagler, J. (1994), 'Scobit: an alternative estimator to logit and probit', *American Journal of Political Science* pp. 230–255.

Nakayama, S. and Chikaraishi, M. (2015), 'Unified closed-form expression of logit and weibit and its extension to a transportation network equilibrium assignment', *Transportation Research Part B: Methodological* **81**, 672–685.

Olsen, S. B., Lundhede, T. H., Jacobsen, J. B. and Thorsen, B. J. (2011), 'Tough and easy choices: testing the influence of utility difference on stated certainty-in-choice in choice experiments', *Environmental and Resource Economics* **49**(4), 491–510.

Paleti, R., Bhat, C. R. and Pendyala, R. M. (2013), 'Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics', *Transportation Research Record* **2382**(1), 162–172.

Patil, P. N., Dubey, S. K., Pinjari, A. R., Cherchi, E., Daziano, R. and Bhat, C. R. (2017), 'Simulation evaluation of emerging estimation techniques for multinomial probit models', *Journal of choice modelling* **23**, 9–20.

Pewsey, A. (2000), 'Problems of inference for azzalini's skewnormal distribution', *Journal of Applied Statistics* **27**(7), 859–870.

Piatek, R. and Gensowski, M. (2017), 'A multinomial probit model with latent factors: Identification and interpretation without a measurement system'.

Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001), 'Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution', *Journal of Computational and Graphical Statistics* **10**(2), 249–276.

Rashidi, T. H., Auld, J. and Mohammadian, A. K. (2012), 'A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection', *Transportation Research Part A: Policy and Practice* **46**(7), 1097–1107.

Richard, J.-F. and Zhang, W. (2007), 'Efficient high-dimensional importance sampling', *Journal of Econometrics* **141**(2), 1385–1411.

Ruud, P. A. (1991), 'Extensions of estimation methods using the EM algorithm', *Journal of Econometrics* **49**(3), 305–341.

Ruud, P. A. (1996), 'Simulation of the multinomial probit model: An analysis of covariance matrix estimation', *Working Paper, Department of Economics, University of California, Berkeley* .

Ryu, E. K. and Boyd, S. P. (2015), 'Extensions of gauss quadrature via linear programming', *Foundations of Computational Mathematics* **15**(4), 953–971.

Sándor, Z. and Train, K. (2004), 'Quasi-random simulation of discrete choice models', *Transportation Research Part B: Methodological* **38**(4), 313–327.

Sarrias, M. and Daziano, R. (2016), 'Multinomial logit models with continuous and discrete individual heterogeneity in r: The gmnl package', *Journal of Statistical Software (Accepted for publication)* .

Schoemaker, P. J. (2013), *Experiments on decisions under risk: The expected utility hypothesis*, Springer Science & Business Media.

Shoup, D. (2005), *The High Cost of Free Parking*, Planners Press, American Planning Association, Chicago.

Smolyak, S. A. (1963), Quadrature and interpolation formulas for tensor products of certain classes of functions, *in* 'Doklady Akademii Nauk', Vol. 148, Russian Academy of Sciences, pp. 1042–1045.

Sohn, K. (2017), 'An expectation-maximization algorithm to estimate the integrated choice and latent variable model', *Transportation Science* **51**(3), 946–967.

Spissu, E., Pinjari, A. R., Pendyala, R. M. and Bhat, C. R. (2009), 'A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel', *Transportation* **36**(4), 403–422.

Stevens, M. R. (2017), 'Does compact development make people drive less?', *Journal of the American Planning Association* **83**(1), 7–18.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01944363.2016.1240044*

Swait, J. (2001), 'A non-compensatory choice model incorporating attribute cutoffs', *Transportation Research Part B: Methodological* **35**(10), 903–928.

Train, K. (2016), 'Mixed logit with a flexible mixing distribution', *Journal of Choice Modelling* **19**, 40–53.

Train, K. E. (2008), 'EM algorithms for nonparametric estimation of mixing distributions', *Journal of Choice Modelling* **1**(1), 40–69.

Train, K. E. (2009), *Discrete choice methods with simulation*, second edn, Cambridge University Press.

Varin, C., Reid, N. and Firth, D. (2011), 'An overview of composite likelihood methods', *Statistica Sinica* pp. 5–42.

Varin, C. and Vidoni, P. (2005), 'A note on composite likelihood inference and model selection', *Biometrika* **92**(3), 519–528.

Vij, A. and Krueger, R. (2017), 'Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions', *Transportation Research Part B: Methodological* **106**, 76–101.

Vijverberg, C.-P. C. and Vijverberg, W. P. (2016), 'Pregibit: a family of binary choice models', *Empirical Economics* **50**(3), 901–932.

von Haefen, R. H. and Domanski, A. (2018), 'Estimation and welfare analysis from mixed logit models with large choice sets', *Journal of Environmental Economics and Management* **90**, 101–118.

Wang, W.-L., Lin, T.-I. and Lachos, V. H. (2018), 'Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails', *Statistical methods in medical research* **27**(1), 48–64.

Xu, X. and Reid, N. (2011), 'On the robustness of maximum composite likelihood estimate', *Journal of Statistical Planning and Inference* **141**(9), 3047–3054.

Yu, J., Goos, P. and Vandebroek, M. (2010), 'Comparing different sampling schemes for approximating the integrals involved in the efficient design of stated choice experiments', *Transportation Research Part B: Methodological* **44**(10), 1268–1289.